Neyman-Pearson Classification under High-Dimensional Settings

Anqi Zhao ANQIZHAO@FAS.HARVARD.EDU
Department of Statistics
Harvard University

Yang Feng YANGFENG@STAT.COLUMBIA.EDU
Department of Statistics
Columbia University

Lie Wang LIEWANG@MATH.MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology

Xin Tong XINT@MARSHALL.USC.EDU
Department of Data Sciences and Operations
Marshall Business School
University of Southern California

Editor:

Abstract

Most existing binary classification methods target on the optimization of the overall classification risk and may fail to serve some real-world applications such as cancer diagnosis, where users are more concerned with the risk of misclassifying one specific class than the other. Neyman-Pearson (NP) paradigm was introduced in this context as a novel statistical framework for handling asymmetric type I/II error priorities. It seeks classifiers with a minimal type II error and a constrained type I error under a user specified level. This article is the first attempt to construct classifiers with guaranteed theoretical performance under the NP paradigm in high-dimensional settings. Based on the fundamental Neyman-Pearson Lemma, we used a plug-in approach to construct NP-type classifiers for Naive Bayes models. The proposed classifiers satisfy the NP oracle inequalities, which are natural NP paradigm counterparts of the oracle inequalities in classical binary classification. Besides their desirable theoretical properties, we also demonstrated their numerical advantages in prioritized error control via both simulation and real data studies.

Keywords: classification, high-dimension, Naive Bayes, Neyman-Pearson (NP) paradigm, NP oracle inequality, plug-in approach, screening

1. Introduction

Classification plays an important role in many aspects of our society. In medical research, identifying pathogenically distinct tumor types is central to advances in cancer treatments (Golub et al., 1999; Alderton, 2014). In cyber security, spam messages and virus make automatic categorical decisions a necessity. Binary classification is arguably the simplest and most important form of classification problems, and can serve as a building block for more complicated applications. We focus our attention on binary classification in this work.

A few common notations are introduced to facilitate our discussion. Let (X, Y) be a random pair where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of features and $Y \in \{0, 1\}$ indicates X's class label. A *classifier* $\phi : \mathcal{X} \to \{0, 1\}$ is a mapping from $\mathcal{X}$ to $\{0, 1\}$ that assigns X to one of the classes. A *classification loss function* is defined to assign a "cost" to each misclassified instance $\phi(X) \neq Y$, and the *classification error* is defined as the expectation of this loss function with respect to the joint distribution of (X, Y). We will focus our discussion on the 0-1 loss function $\mathbb{1}\{\phi(X) \neq Y\}$ throughout the paper, where $\mathbb{1}(\cdot)$ denotes the indicator function. Denote by $\mathbb{P}$ and $\mathbb{E}$ the generic probability distribution and expectation, whose meaning depends on specific contexts. The classification error is $R(\phi) = \mathbb{E}\mathbb{1}\{\phi(X) \neq Y\} = \mathbb{P}\{\phi(X) \neq Y\}$. The law of total probability allows us to decompose it into a weighted average of type I error $R_0(\phi) = \mathbb{P}\{\phi(X) \neq Y | Y = 0\}$ and type II error $R_1(\phi) = \mathbb{P}\{\phi(X) \neq Y | Y = 1\}$ as

$$R(\phi) \,=\, \mathbb{P}(Y = 0)R_0(\phi) + \mathbb{P}(Y = 1)R_1(\phi)\,. \tag{1.1}$$

With the advent of high-throughput technologies, classification tasks have experienced an exponential growth in the feature dimensions throughout the past decade. The fundamental challenge of "high dimension, low sample size" has motivated the development of a plethora of classification algorithms for various applications. While dependencies among features are usually considered a crucial characteristic of the data (Ackermann and Strimmer, 2009), and can effectively reduce classification errors under suitable models and relative data abundance (Shao et al., 2011; Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012; Witten and Tibshirani, 2012), independence rules, with their superb scalability, become a rule of thumb when the feature dimension grows faster than the sample size (Hastie et al., 2009; James et al., 2013). Despite Naive Bayes models' reputation of being "simplistic" by ignoring all dependency structure among features, they lead to simple classifiers that have proven worthy on high-dimensional data with remarkably good performances in numerous real-life applications. Taking the classical model setting of two-class Gaussian with a common covariance matrix, Bickel and Levina (2004) showed the superior performance of Naive Bayes models over (naive implementation of) the Fisher linear discriminant rule under broad conditions in high-dimensional settings. Fan and Fan (2008) further established the necessity of feature selection for high-dimensional classification problems by showing that even independence rules can be as poor as random guessing due to noise accumulation. Featuring both independence rule and feature selection, the (sparse) Naive Bayes model remains a good choice for classification when the sample size is *fairly limited*.

1.1 Asymmetrical priorities on errors

Most existing binary classification methods target on the optimization of the overall risk (1.1) and may fail to serve the purpose when users' relative priorities over type I/II errors differ significantly from those implied by the marginal probabilities of the two classes. A representative example of such scenario is the diagnosis of serious disease. Let 1 code the healthy class and 0 code the diseased class. Given that usually

$$\mathbb{P}(Y = 1) \gg \mathbb{P}(Y = 0)\,,$$

minimizing the overall risk (1.1) might yield classifiers with small overall risk R (as a result of small R_1) yet large R_0 — a situation quite undesirable in practice given flagging a healthy

case incurs only extra cost of additional tests while failing to detect the disease endangers a life.

The neuroblastoma dataset introduced by Oberthuer et al. (2006) provides a perfect illustration of such intuition. The dataset contains gene expression profiles on $d = 10707$ genes from 246 patients in a German neuroblastoma trial, among which 56 are high-risk (labeled as 0) and 190 are low-risk (labeled as 1). We randomly selected 41 '0's and 123 '1's as our training sample (such that the proportion of '0's is about the same as that in the entire dataset), and tested the resulting classifiers on the rest 15 '0's and 67 '1's. The average error rates of PSN^2 (to be proposed; implemented here at significance level 0.05), Gaussian Naive Bayes (nb), penalized logistic regression (pen-log), and Support Vector Machine (svm) over 1000 random splits are summarized in Table 1. All procedures except PSN^2 led to high type I errors, and are thus considered unsatisfactory given the more severe consequences of missing a diseased instance than vice versa.

Table 1: Average error rates over 1000 random splits for neuroblastoma dataset.

Error Type	PSN^2	nb	pen-log	svm
type I (0 as 1)	.038	.308	.529	.603
type II (1 as 0)	.761	.150	.103	.573

One existing solution to asymmetric error control is *cost-sensitive learning*, which assigns two different costs as weights of the type I/II errors (Elkan, 2001; Zadrozny et al., 2003). Despite many merits and practical values of this framework, limitations arise in applications when there is no consensus over how much costs to be assigned to each class, or more fundamentally, whether it is morally acceptable to assign costs in the first place. Also, when users have a specific target for type I/II error control, cost-sensitive learning does not fit. Other methods aiming for small type I error include the Asymmetric Support Vector Machine (Wu et al., 2008), and the p-value for classification (Dümbgen et al., 2008). However, the former has no theoretical guarantee on errors, while the latter treats all classes as of equal importance.

1.2 Neyman-Pearson (NP) paradigm and NP oracle inequalities

Neyman-Pearson (NP) paradigm was introduced as a novel statistical framework for targeted type I/II error control. Assume type I error R_0 as the prioritized error type, this paradigm seeks to control R_0 under a user specified level α with R_1 as small as possible. The *oracle* is thus

$$\phi^* \in \operatorname*{argmin}_{R_0(\phi)\leq\alpha} R_1(\phi)\,, \tag{1.2}$$

where the *significance level* α reflects the level of conservativeness towards type I error. Given ϕ^* is unattainable in the learning paradigm, the best within our capability is to construct a data dependent classifier $\hat{\phi}$ that mimics it.

Despite its practical importance, NP classification has not received much attention in the statistics and machine learning communities. Cannon et al. (2002) initiated the theoretical treatment of NP classification. Under the same framework, Scott (2005) and Scott and

Nowak (2005) derived several results for traditional statistical learning such as PAC bounds or oracle inequalities. By combining type I and type II errors in sensible ways, Scott (2007) proposed a performance measure for NP classification. More recently, Blanchard et al. (2010) developed a general solution to semi-supervised novelty detection by reducing it to NP classification. Other related works include Casasent and Chen (2003) and Han et al. (2008). A common issue with methods in this line of literature is that they all follow an empirical risk minimization (ERM) approach, and use some forms of relaxed empirical type I error constraint in the optimization program. As a result, all type I errors can only be proven to satisfy some relaxed upper bound. Take the framework set up by Cannon et al. (2002) for example. Given $\varepsilon_0 > 0$, they proposed the program

$$\min_{\phi \in \mathcal{H}, \hat{R}_0(\phi) \leq \alpha + \varepsilon_0/2} \hat{R}_1(\phi)\,,$$

where $\mathcal{H}$ is a set of classifiers with finite Vapnik-Chervonenkis dimension, and $\hat{R}_0$, $\hat{R}_1$ are the empirical type I and type II errors respectively. It is shown that with high probability, the solution $\hat{\phi}$ to the above program satisfies simultaneously: i) the type I error $R_0(\hat{\phi})$ is bounded from above by $\alpha + \varepsilon_0$, and ii) the type II error $R_1(\hat{\phi})$ is bounded from above by $R_1(\phi^*) + \varepsilon_1$ for some $\varepsilon_1 > 0$.

Rigollet and Tong (2011) is a significant departure from the previous NP classification literature. This paper argues that a good classifier $\hat{\phi}$ under the NP paradigm should respect the chosen significance level α, rather than some relaxation of it. More precisely, two **NP oracle inequalities** should be satisfied simultaneously with high probability:

(I) the type I error constraint is respected, i.e., $R_0(\hat{\phi}) \leq \alpha$.

(II) the excess type II error $R_1(\hat{\phi}) - R_1(\phi^*)$ diminishes with explicit rates (w.r.t. sample size).

Recall that, for a classifier $\hat{h}$, the classical oracle inequality insists that with high probability

$$\text{the excess risk } R(\hat{h}) - R(h^*) \text{ diminishes with explicit rates,} \tag{1.3}$$

where $h^*(x) = \mathbb{I}(\eta(x) \geq 1/2)$ is the Bayes classifier, in which $\eta(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$ is the regression function of Y on X (see Koltchinskii (2008) and references within). The two NP oracle inequalities defined above can be thought of as a generalization of (1.3) that provides a novel characterization of classifiers' theoretical performances under the NP paradigm.

Using a more stringent empirical type I error constraint (than the level α), Rigollet and Tong (2011) established NP oracle inequalities for its proposed classifiers under convex loss functions (as opposed to the indicator loss). They also proved an interesting negative result: under the binary loss, ERM approaches (convexification or not) cannot guarantee diminishing excess type II error as long as one insists type I error of the proposed classifier be bounded from above by α with high probability. This negative result motivated a plug-in approach to NP classification in Tong (2013).

1.3 Plug-in approaches

Plug-in methods in classical binary classification have been well studied in the literature, where the usual plug-in target is the Bayes classifier $\mathbb{I}(\eta(x) \geq 1/2)$. Earlier works gave rise to pessimism of the plug-in approach to classification. For example, under certain assumptions, Yang (1999) showed plug-in estimators cannot achieve excess risk with rates faster than $O(1/\sqrt{n})$, while direct methods can achieve rates up to $O(1/n)$ under *margin assumption* (Mammen and Tsybakov, 1999; Tsybakov, 2004; Tsybakov and van de Geer, 2005; Tarigan and van de Geer, 2006). However, it was shown in Audibert and Tsybakov (2007) that plug-in classifiers $\mathbb{I}(\hat{\eta}_n \geq 1/2)$ based on local polynomial estimators can achieve rates faster than $O(1/n)$, with a smoothness condition on η and the margin assumption.

The oracle classifier under the NP paradigm arises from its close connection to the Neyman-Pearson Lemma in statistical hypothesis testing. Hypothesis testing bears strong resemblance to binary classification if we assume the following model. Let P_1 and P_0 be two *known* probability distributions on $\mathcal{X} \subset \mathbb{R}^d$. Assume that $Y \sim \text{Bern}(\zeta)$ for some $\zeta \in (0,1)$, and the conditional distribution of X given Y is P_Y. Given such a model, the goal of statistical hypothesis testing is to determine if we should reject the null hypothesis that X was generated from P_0. To this end, we construct a randomized test $\phi : \mathcal{X} \to [0,1]$ that rejects the null with probability $\phi(X)$. Two types of errors arise: type I error occurs when P_0 is rejected yet $X \sim P_0$, and type II error occurs when P_0 is not rejected yet $X \sim P_1$. The Neyman-Pearson paradigm in hypothesis testing amounts to choosing ϕ that solves the following constrained optimization problem

$$\text{maximize } \mathbb{E}[\phi(X)|Y=1]\,, \text{ subject to } \mathbb{E}[\phi(X)|Y=0] \leq \alpha\,,$$

where $\alpha \in (0,1)$ is the significance level of the test. A solution to this constrained optimization problem is called *a most powerful test* of level α. The Neyman-Pearson Lemma gives mild sufficient conditions for the existence of such a test.

Lemma 1.1 (Neyman-Pearson Lemma)**.** *Let P_1 and P_0 be two probability measures with densities p and q respectively, and denote the density ratio as $r(x) = p(x)/q(x)$. For a given significance level α, let C_α be such that $P_0\{r(X) > C_\alpha\} \leq \alpha$ and $P_0\{r(X) \geq C_\alpha\} \geq \alpha$. Then, the most powerful test of level α is*

$$\phi^*(X) = \begin{cases} 1 & \textit{if } \; r(X) > C_\alpha\,, \\ 0 & \textit{if } \; r(X) < C_\alpha\,, \\ \frac{\alpha - P_0\{r(X)>C_\alpha\}}{P_0\{r(X)=C_\alpha\}} & \textit{if } \; r(X) = C_\alpha\,. \end{cases}$$

Under mild continuity assumption, we take the NP *oracle*

$$\phi^*(x) \;=\; \phi^*_\alpha(x) \;=\; \mathbb{I}\{p(x)/q(x) \geq C_\alpha\} \;=\; \mathbb{I}\{r(x) \geq C_\alpha\}\,. \tag{1.4}$$

as our plug-in target for NP classification. With kernel density estimates $\hat{p}$, $\hat{q}$, and a proper estimate of the threshold level $\widehat{C}_\alpha$, Tong (2013) constructed a plug-in classifier $\mathbb{I}\{\hat{p}(x)/\hat{q}(x) \geq \widehat{C}_\alpha\}$ that satisfies both NP oracle inequalities with high probability when the dimensionality is small, leaving the high-dimensional case an unchartered territory.

1.4 Contribution

In the big data era, NP classification framework faces the same curse of dimensionality as its classical counterpart. Despite its wide potential applications, this paper is the *first attempt* to construct performance-guaranteed classifiers under the NP paradigm in high-dimensional settings. Based on the Neyman-Pearson Lemma, we employ Naive Bayes models and propose a computationally feasible plug-in approach to construct classifiers that satisfy the NP oracle inequalities. We also improve the *detection condition*, a critical theoretical assumption first introduced in Tong (2013), for effective threshold level estimation that grounds the good NP properties of these classifiers. Necessity of the new detection condition is also discussed. Note that classifiers proposed in this work are not straightforward extensions of Tong (2013): kernel density estimation is now applied in combination with feature selection, and the threshold level is estimated in a more precise way by order statistics that require only moderate sample size — while Tong (2013) resorted to the Vapnik-Chervonenkis theory and required sample size much bigger than what is available in most high-dimensional applications.

The rest of the paper is organized as follows. Two screening based plug-in NP-type classifiers are presented in Section 2, where theoretical properties are also discussed. Performance of the proposed classifiers is demonstrated in Section 3 by both simulation studies and real data analysis. We conclude in Section 4 with a short discussion. The technical proofs are relegated to the Appendix.

2. Methods

In this section, we first introduce several notations and definitions, with a focus on the *detection condition*. Then we present the plug-in procedure, together with its theoretical properties.

2.1 Notations and definitions

We introduce here several notations adapted from Audibert and Tsybakov (2007). For $\beta > 0$, denote by $\lfloor \beta \rfloor$ the largest integer strictly less than β. For any $x, x' \in \mathbb{R}$ and any $\lfloor \beta \rfloor$ times continuously differentiable real-valued function $g(\cdot)$ on $\mathbb{R}$, we denote by g_x its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point x. For $L > 0$, the $(\beta, L, [-1, 1])$-Hölder class of functions, denoted by $\Sigma(\beta, L, [-1, 1])$, is the set of functions $g : [-1, 1] \to \mathbb{R}$ that are $\lfloor \beta \rfloor$ times continuously differentiable and satisfy, for any $x, x' \in [-1, 1]$, the inequality $|g(x') - g_x(x')| \leq L|x - x'|^\beta$. The $(\beta, L, [-1, 1])$-Hölder class of density is defined as

$$\mathcal{P}_\Sigma(\beta, L, [-1, 1]) = \left\{ f \,:\, f \geq 0, \int f = 1, f \in \Sigma(\beta, L, [-1, 1]) \right\} .$$

We will use β-valid kernels (kernels of order β, Tsybakov (2009)) for all the kernel estimation throughout the theoretical discussion, the definition of which is as follows.

Definition 2.1. *Let $K(\cdot)$ be a real-valued function on $\mathbb{R}$ with support $[-1, 1]$. The function $K(\cdot)$ is a β-valid kernel if it satisfies $\int K = 1$, $\int |K|^v < \infty$ for any $v \geq 1$, $\int |t|^\beta |K(t)| dt < \infty$, and in the case $\lfloor \beta \rfloor \geq 1$, it satisfies $\int t^l K(t) dt = 0$ for any $l \in \mathbb{N}$ such that $1 \leq l \leq \lfloor \beta \rfloor$.*

We assume that all the β-valid kernels considered in the theoretical part of this paper are constructed from Legendre polynomials, and are thus Lipschitz and bounded, satisfying the kernel conditions for the important technical Lemma A.6.

Definition 2.2 (margin assumption)**.** *A function $f(\cdot)$ is said to satisfy margin assumption of order $\bar{\gamma}$ with respect to probability distribution P at the level C^* if there exists a positive constant M_0, such that for any $\delta \geq 0$,*

$$P\{|f(X) - C^*| \leq \delta\} \leq M_0 \delta^{\bar{\gamma}} .$$

This assumption was first introduced in Polonik (1995). In the classical binary classification framework, Mammen and Tsybakov (1999) proposed a similar condition named "margin condition" by requiring most data to be away from the optimal decision boundary. In the classical classification paradigm, definition 2.2 reduces to the "margin condition" by taking $f = \eta$ and $C^* = 1/2$, with $\{x : |f(x) - C^*| = 0\} = \{x : \eta(x) = 1/2\}$ giving the decision boundary of the Bayes classifier. On the other hand, unlike the classical paradigm where the optimal threshold level is known and does not need an estimate, the optimal threshold level C_α in the NP paradigm is unknown and needs to be estimated, suggesting the necessity of having sufficient data around the decision boundary to detect it well. This concern motivated the following condition improved from Tong (2013).

Definition 2.3 (detection condition)**.** *A function $f(\cdot)$ is said to satisfy detection condition of order $\underline{\gamma}$ with respect to P (i.e., $X \sim P$) at level (C^*, δ^*) if there exists a positive constant M_1, such that for any $\delta \in (0, \delta^*)$,*

$$P\{C^* \leq f(X) \leq C^* + \delta\} \geq M_1 \delta^{\underline{\gamma}} .$$

A detection condition works as an opposite force to the margin assumption, and is basically an assumption on the lower bound of probability. Though we take here a power function as the lower bound, so that it is simple and aesthetically similar to the margin assumption, any increasing $u(\cdot)$ on R^+ with $\lim_{x\to 0+} g(x) = 0$ should be able to serve the purpose. The version of detection condition we would use to establish the NP inequalities for the (to be) proposed classifiers takes $f = r$, $C^* = C_\alpha$, and $P = P_0$ (recall that P_0 is the conditional distribution of X given $Y = 0$).

Now we argue why such a condition is *necessary* to achieve the NP oracle inequalities. Consider the simpler case where the density ratio r is known, and we only need a proper estimate of the threshold level $\widehat{C}_\alpha$. If there is nothing like the detection condition (Definition 2.3 involves a power function, but the idea is just to have any kind of lower bound), we would have, for some $\delta > 0$,

$$P_0\{C_\alpha \leq r(X) \leq C_\alpha + \delta\} = 0 . \tag{2.1}$$

In getting the threshold estimate $\widehat{C}_\alpha$ of $\hat{\phi}(x) = \mathbb{I}\{r(x) \geq \widehat{C}_\alpha\}$, we can not distinguish any threshold level between C_α and $C_\alpha + \delta$. In particular, it is possible that

$$\widehat{C}_\alpha > C_\alpha + \delta/2 .$$

But then the excess type II error is bounded from below as follows

$$R_1(\hat{\phi}) - R_1(\phi^*) = P_1\{C_\alpha < r(X) < \widehat{C}_\alpha\} > P_1\{C_\alpha < r(X) < C_\alpha + \delta/2\}\,,$$

where the last quantity can be positive. Therefore, the second NP oracle inequality (diminishing excess type II error) does not hold for $\hat{\phi}$. Since some detection condition is necessary in this simpler case, it is certainly necessary in our real setup.

Note that Definition 2.3 is a significant improvement of the detection condition formulated in Tong (2013), which requires

$$P\{C^* - \delta \leq f(X) \leq C^*\} \wedge P\{C^* \leq f(X) \leq C^* + \delta\} \;\geq\; M_1 \delta^{\underline{\gamma}}\,.$$

We are able to drop the lower bound for the first piece due to an improved layout of the proofs. Intuitively, our new detection condition ensures an upper bound on $\widehat{C}_\alpha$. But we do not need an extra condition to get a lower bound of $\widehat{C}_\alpha$, because of the type I error bound requirement (see the proof of Proposition 2.4 for details).

2.2 Neyman-Pearson plug-in procedure

Suppose the sampling scheme is fixed as follows.

Assumption 1. *Assume the training sample contains n i.i.d. observations $\mathcal{S}^1 = \{U_1, \cdots, U_n\}$ from class 1 with density p, and m i.i.d. observations $\mathcal{S}^0 = \{V_1, \cdots, V_m\}$ from class 0 with density q. Given fixed n_1, n_2, m_1, m_2 and m_3 such that $n_1 + n_2 = n$, $m_1 + m_2 + m_3 = m$, we further decompose $\mathcal{S}^1$ and $\mathcal{S}^0$ into independent subsamples as: $\mathcal{S}^1 = \mathcal{S}^1_1 \cup \mathcal{S}^1_2$, and $\mathcal{S}^0 = \mathcal{S}^0_1 \cup \mathcal{S}^0_2 \cup \mathcal{S}^0_3$, where $|\mathcal{S}^1_1| = n_1$, $|\mathcal{S}^1_2| = n_2$, $|\mathcal{S}^0_1| = m_1$, $|\mathcal{S}^0_2| = m_2$, $|\mathcal{S}^0_3| = m_3$.*

The sample splitting idea has been considered in the literature, such as in Meinshausen and Bühlmann (2010) and Robins et al. (2006). Given these samples, we introduce the following plug-in procedure.

Definition 2.4. Neyman-Pearson plug-in procedure

<u>*Step 1*</u> *Use $\mathcal{S}^1_1$, $\mathcal{S}^1_2$, $\mathcal{S}^0_1$, and $\mathcal{S}^0_2$ to construct a density ratio estimate $\hat{r}$. The specific use of each subsample will be introduced in Section 2.4.*

<u>*Step 2*</u> *Given $\hat{r}$, choose a threshold estimate $\widehat{C}_\alpha$ from the set $\hat{r}(\mathcal{S}^0_3) = \{\hat{r}(V_{i+m_1+m_2})\}_{i=1}^{m_3}$.*

Denote by $\hat{r}_{(k)}(\mathcal{S}^0_3)$ the k-th order statistic of $\hat{r}(\mathcal{S}^0_3)$, $k \in \{1, \cdots, m_3\}$. The corresponding plug-in classifier by setting $\widehat{C}_\alpha = \hat{r}_{(k)}(\mathcal{S}^0_3)$ is

$$\hat{\phi}_k(x) = \mathbb{1}\{\hat{r}(x) \geq \hat{r}_{(k)}(\mathcal{S}^0_3)\}\,. \tag{2.2}$$

A generic procedure for choosing the optimal k will be given in Section 2.3.

2.3 Threshold estimate $\widehat{C}_\alpha$

For any arbitrary density ratio estimate $\hat{r}$, we employ a proper order statistic $\hat{r}_{(k)}(\mathcal{S}^0_3)$ to estimate the threshold C_α, and establish a probabilistic upper bound for the type I error of $\hat{\phi}_k$ for each $k \in \{1, \cdots, m_3\}$.

Proposition 2.1. *For any arbitrary density ratio estimate $\hat{r}$, let $\hat{\phi}_k(x) = \mathbb{1}\{\hat{r}(x) \geq \hat{r}_{(k)}(\mathcal{S}_3^0)\}$. It holds for any $\delta \in (0,1)$ and $k \in \{1, \cdots, m_3\}$ that*

$$\mathbb{P}\{R_0(\hat{\phi}_k) > \delta\} \leq \text{Beta.cdf}_{k,\, m_3+1-k}\,(1-\delta)\ , \tag{2.3}$$

where $\text{Beta.cdf}_{k,\, m_3+1-k}(\cdot)$ is the CDF *of $\text{Beta}(k,\, m_3+1-k)$. The inequality becomes equality when $F_{0,\hat{r}}(t) = P_0\{\hat{r}(X) \leq t\}$ is continuous almost surely.*

In view of the above proposition, a sufficient condition for the classifier $\hat{\phi}_k$ to satisfy NP Oracle Inequality (I) at tolerance level $\delta_3 \in (0,1)$ is thus

$$\text{Beta.cdf}_{k,\, m_3+1-k}\,(1-\alpha) \leq \delta_3\,. \tag{2.4}$$

Despite the potential tightness of (2.3), we are not able to derive an explicit formula for the minimum k that satisfies (2.4). To get an explicit choice for k, we resort to concentration inequalities for an alternative.

Proposition 2.2. *For any arbitrary density ratio estimate $\hat{r}$, let $\hat{\phi}_k(x) = \mathbb{1}\{\hat{r}(x) \geq \hat{r}_{(k)}(\mathcal{S}_3^0)\}$. It holds for any $\delta_3 \in (0,1)$ and $k \in \{1, \cdots, m_3\}$ that*

$$\mathbb{P}\{R_0(\hat{\phi}_k) > g(\delta_3, m_3, k)\} \leq \delta_3\,, \tag{2.5}$$

where

$$g(\delta_3, m_3, k) = \frac{m_3+1-k}{m_3+1} + \sqrt{\frac{k(m_3+1-k)}{\delta_3(m_3+2)(m_3+1)^2}}\,. \tag{2.6}$$

Let $\mathcal{K} = \mathcal{K}(\alpha, \delta_3, m_3) = \{k \in \{1, \cdots, m_3\} : g(\delta_3, m_3, k) \leq \alpha\}$. Proposition 2.2 implies that $k \in \mathcal{K}(\alpha, \delta_3, m_3)$ is a sufficient condition for the classifier $\hat{\phi}_k$ to satisfy NP Oracle Inequality (I). The next step is to characterize $\mathcal{K}$ and choose some $k \in \mathcal{K}$, so that $\hat{\phi}_k$ has small excess type II error. Clearly, we would like to find the smallest element in $\mathcal{K}$.

Proposition 2.3. *The minimum $k \in \{1, \cdots, m_3+1\}$ that satisfies $g(\delta_3, m_3, k) \leq \alpha$ is*

$$k_{\min}(\alpha, \delta_3, m_3) = \lceil (m_3+1) A_{\alpha,\delta_3}(m_3) \rceil\,, \tag{2.7}$$

where $\lceil z \rceil$ denotes the smallest integer larger than or equal to z, and

$$A_{\alpha,\delta_3}(m_3) = \frac{1 + 2\delta_3(m_3+2)(1-\alpha) + \sqrt{1 + 4\delta_3(1-\alpha)\alpha(m_3+2)}}{2\left\{\delta_3(m_3+2)+1\right\}}\,.$$

Moreover,

1. *$A_{\alpha,\delta_3}(m_3) \in (1-\alpha, 1)$.*
2. *$\hat{r}_{(k_{\min}(\alpha,\delta_3,m_3))}(\mathcal{S}_3^0)$ is asymptotically the empirical $(1-\alpha)$-th quantile of $F_{0,\hat{r}}$ in the sense that*

$$\lim_{m_3\to\infty} \frac{k_{\min}(\alpha, \delta_3, m_3)}{m_3} = \lim_{m_3\to\infty} A_{\alpha,\delta_3}(m_3) = 1-\alpha\,.$$

3. For any $m_3 \geq 4/(\alpha\delta_3)$, we have $k_{\min}(\alpha, \delta_3, m_3) \leq m_3$, and thus

$$\mathcal{K}(\alpha, \delta_3, m_3) \;=\; \{k_{\min}(\alpha, \delta_3, m_3), k_{\min}(\alpha, \delta_3, m_3) + 1, \ldots, m_3\}\,.$$

Introduce shorthand notations $k_{\min} = k_{\min}(\alpha, \delta_3, m_3)$, $\hat{r}_{(k)} = \hat{r}_{(k)}(\mathcal{S}_3^0)$, and $\widehat{C}_\alpha = \hat{r}_{(\min\{k_{\min}, m_3\})}$. We will take

$$\hat{\phi}(x) \;=\; \mathbb{I}\{\hat{r}(x) \geq \widehat{C}_\alpha\} \;=\; \begin{cases} \mathbb{I}\{\hat{r}(x) \geq \hat{r}_{(k_{\min})}\}\,, & \text{if} \;\; k_{\min} \leq m_3\,, \\ \mathbb{I}\{\hat{r}(x) \geq \hat{r}_{(m_3)}\}\,, & \text{if} \;\; k_{\min} = m_3 + 1 \end{cases} \tag{2.8}$$

as the default NP plug-in classifier for any arbitrary $\hat{r}$. An alternative threshold estimate that also guarantees type I error bound is derived in the Appendix C. Assume $m_3 \geq 4/(\alpha\delta_3)$ for the rest of the theoretical discussion. It follows from Proposition 2.3 that $k_{\min} \leq m_3$, and thus $\widehat{C}_\alpha = \hat{r}_{(k_{\min})}$, $\hat{\phi} = \hat{\phi}_{(k_{\min})}$ with guaranteed type I error control.

Remark 2.1. *Note that $\lim_{m_3 \to \infty} k_{\min}/\lceil m_3(1-\alpha)\rceil = 1$. Thus, choosing the $k_{\min}$-th order statistic of $\hat{r}(\mathcal{S}_3^0)$ as the threshold can be viewed as a modification to the classical approach of estimating the $1-\alpha$ quantile of $F_{0,\hat{r}}$ by the $\lceil m_3(1-\alpha)\rceil$-th order statistic of $\hat{r}(\mathcal{S}_3^0)$. Recall that the oracle C_α is actually the $1-\alpha$ quantile of distribution $F_{0,r}$, so the intuition is that $\widehat{C}_\alpha$ is asymptotically (when $m_3 \to \infty$) equivalent to the $1-\alpha$ quantile of $F_{0,\hat{r}}$, which in turn converges (when $n_1, n_2, m_1, m_2 \to \infty$) to C_α as the $1-\alpha$ quantile of $F_{0,r}$ under moderate conditions.*

Lemma 2.1. *Let $\alpha, \delta_3 \in (0,1)$. In addition to Assumption 1, suppose $\hat{r}$ be such that $F_{0,\hat{r}}$ is continuous almost surely. Then for any $\delta_4 \in (0,1)$ and $m_3 \geq 4/(\alpha\delta_3)$, the distance between $R_0(\hat{\phi})$ ($\hat{\phi}$ as defined in (2.8)) and $R_0(\phi^*)$ can be bounded as*

$$\mathbb{P}\{|R_0(\hat{\phi}) - R_0(\phi^*)| > \xi_{\alpha,\delta_3,m_3}(\delta_4)\} \;\leq\; \delta_4\,,$$

where

$$\xi_{\alpha,\delta_3,m_3}(\delta_4) = \sqrt{\frac{k_{\min}(m_3 + 1 - k_{\min})}{(m_3+2)(m_3+1)^2\delta_4}} + A_{\alpha,\delta_3}(m_3) - (1-\alpha) + \frac{1}{m_3+1}\,. \tag{2.9}$$

If $m_3 \geq \max(\delta_3^{-2}, \delta_4^{-2})$, we have $\xi_{\alpha,\delta_3,m_3}(\delta_4) \leq (5/2)m_3^{-1/4}$.

Proposition 2.4. *Let $\alpha, \delta_3, \delta_4 \in (0,1)$. In addition to assumptions of Lemma 2.1, assume that the density ratio r satisfies the margin assumption of order $\bar{\gamma}$ at level C_α (with constant M_0) and detection condition of order $\underline{\gamma}$ at level (C_α, δ^*) (with constant M_1), both with respect to distribution P_0. If $m_3 \geq \max\{4/(\alpha\delta_3), \delta_3^{-2}, \delta_4^{-2}, (\frac{2}{5}M_1{\delta^*}^{\underline{\gamma}})^{-4}\}$, the excess type II error of the classifier $\hat{\phi}$ defined in (2.8) satisfies with probability at least $1 - \delta_3 - \delta_4$,*

$$\begin{aligned} &R_1(\hat{\phi}) - R_1(\phi^*) \\ &\leq\; 2M_0\left[\left\{\frac{|R_0(\hat{\phi}) - R_0(\phi^*)|}{M_1}\right\}^{1/\underline{\gamma}} + 2\|\hat{r} - r\|_\infty\right]^{1+\bar{\gamma}} + C_\alpha|R_0(\hat{\phi}) - R_0(\phi^*)| \\ &\leq\; 2M_0\left[\left(\frac{2}{5}m_3^{1/4}M_1\right)^{-1/\underline{\gamma}} + 2\|\hat{r} - r\|_\infty\right]^{1+\bar{\gamma}} + C_\alpha\left(\frac{2}{5}m_3^{1/4}\right)^{-1}\,. \end{aligned}$$

Given the above proposition, we can control the excess type II error as long as the uniform deviation of density ratio estimate $\|\hat{r} - r\|_\infty$ is controlled. In the following subsection, we will introduce estimates $\hat{r}$ and provide bounds for $\|\hat{r} - r\|_\infty$.

2.4 Density ratio estimate $\hat{r}$

Denote the marginal densities of class 1 and 0 as p_j and q_j ($j = 1, \cdots, d$) respectively, Naive Bayes models for the density ratio take the form

$$r(x) = \prod_{j=1}^{d} \frac{p_j(x_j)}{q_j(x_j)}, \quad \text{where } x_j \text{ is the } j\text{-th component of } x\,.$$

The subsamples $\mathcal{S}_1^1 = \{U_i\}_{i=1}^{n_1}$, $\mathcal{S}_2^1 = \{U_{i+n_1}\}_{i=1}^{n_2}$, $\mathcal{S}_1^0 = \{V_i\}_{i=1}^{m_1}$ and $\mathcal{S}_2^0 = \{V_{i+m_1}\}_{i=1}^{m_2}$ are used to construct (nonparametric/parametric) estimates of p_j and q_j for $j = 1, \cdots, d$.

Nonparametric estimate of the density ratio. For marginal densities p_j and q_j, we apply kernel estimates $\hat{p}_j(x_j) = \{(n_1+n_2)h_1\}^{-1}\sum_{i=1}^{n_1+n_2} K\left(\frac{U_{i,j}-x_j}{h_1}\right)$, and $\hat{q}_j(x_j) = \{(m_1+m_2)h_0\}^{-1}\sum_{i=1}^{m_1+m_2} K\left(\frac{V_{i,j}-x_j}{h_0}\right)$, where $K(\cdot)$ is the kernel function, h_1, h_0 are the bandwidths, and $V_{i,j}$ and $U_{i,j}$ denote the j-th component of V_i and U_i respectively. The resulting nonparametric estimate is

$$\hat{r}_{\mathrm{N}}(x) = \prod_{j=1}^{d} \frac{\hat{p}_j(x_j)}{\hat{q}_j(x_j)}\,. \tag{2.10}$$

Parametric estimate of the density ratio. Assume the two-class Gaussian model $X|Y=0 \sim \mathcal{N}(\mu^0, \Sigma)$ and $X|Y=1 \sim \mathcal{N}(\mu^1, \Sigma)$, where $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_d^2)$. We estimate μ^0, μ^1 and Σ using their sample versions $\hat{\mu}^0$, $\hat{\mu}^1$ and $\hat{\Sigma}$. Under this model, the density ratio function is given by

$$r_{\mathrm{P}}(x) \;=\; \exp\left\{\left(\mu^1 - \mu^0\right)' \Sigma^{-1} x + \frac{1}{2}(\mu^0)'\Sigma^{-1}\mu^0 - \frac{1}{2}(\mu^1)'\Sigma^{-1}\mu^1\right\},$$

and the corresponding parametric estimate is

$$\hat{r}_{\mathrm{P}}(x) \;=\; \exp\left\{\left(\hat{\mu}^1 - \hat{\mu}^0\right)' \hat{\Sigma}^{-1} x + \frac{1}{2}(\hat{\mu}^0)'\hat{\Sigma}^{-1}\hat{\mu}^0 - \frac{1}{2}(\hat{\mu}^1)'\hat{\Sigma}^{-1}\hat{\mu}^1\right\}. \tag{2.11}$$

2.5 Screening-based density ratio estimate and plug-in procedures

For "high dimension, low sample size" applications, complex models that take into account all features usually fail; even Naive Bayes models that ignore feature dependency might lead to poor performance due to noise accumulation (Fan and Fan, 2008). A common solution in these scenarios is to first study marginal relations between the response and each of the features (Fan and Lv, 2008; Li et al., 2012). By selecting the most important individual features, we greatly reduce the model size, and other models can be applied after this screening step. We now introduce screening based variants of $\hat{r}_{\mathrm{N}}$ and $\hat{r}_{\mathrm{P}}$. Let F_j^0 and F_j^1 denote the CDFs of q_j and p_j respectively, for $j = 1, \cdots, d$. Step 1 of Procedure 2.4

introduced in Section 2.1 is now decomposed into a screening substep and an estimation substep.

Nonparametric Screening-based NP Naive Bayes (NSN^2) classifier

Step 1.1 Select features using $\mathcal{S}_1^0$ and $\mathcal{S}_1^1$ as follows:

$$\widehat{\mathcal{A}}_\tau = \left\{1 \leq j \leq d : \|\hat{F}_j^0 - \hat{F}_j^1\|_\infty \geq \tau\right\}, \tag{2.12}$$

where $\tau > 0$ is some threshold level, and

$$\hat{F}_j^0(x_j) = \frac{1}{m_1}\sum_{i=1}^{m_1} \mathbb{1}(V_{i,j} \leq x_j), \quad \hat{F}_j^1(x_j) = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}(U_{i,j} \leq x_j) \tag{2.13}$$

are the empirical CDFs.

Step 1.2 Use $\mathcal{S}_2^0$ and $\mathcal{S}_2^1$ to construct kernel estimates of q_j and p_j for $j \in \widehat{\mathcal{A}}_\tau$. The density ratio estimate is given by

$$\hat{r}_{\mathrm{N}}^S(x) = \prod_{j \in \widehat{\mathcal{A}}_\tau} \frac{\hat{p}_j(x_j)}{\hat{q}_j(x_j)}.$$

Step 2 Given $\hat{r}_{\mathrm{N}}^S$, use $\mathcal{S}_3^0$ to get a threshold estimate $(\hat{r}_{\mathrm{N}}^S)_{(k_{\min})}$ as in (2.8).

The resulting NSN^2 classifier is

$$\hat{\phi}_{\mathrm{NSN}^2}(x) = \mathbb{1}\left\{\hat{r}_{\mathrm{N}}^S(x) \geq (\hat{r}_{\mathrm{N}}^S)_{(k_{\min})}\right\}. \tag{2.14}$$

Parametric Screening-based NP Naive Bayes (PSN^2) classifier

The PSN^2 procedure is similar to NSN^2, except the following two differences. In Step 1.1, features are now selected based on t-statistics ($\widetilde{\mathcal{A}}_\tau$ represent the index set of the selected features). In Step 1.2, p_j, q_j for $j \in \widetilde{\mathcal{A}}_\tau$ follow two-class Gaussian model, and the resulting parametric screening-based density ratio estimate is

$$\hat{r}_{\mathrm{P}}^S(x) = \prod_{j \in \widetilde{\mathcal{A}}_\tau} \frac{\tilde{p}_j(x_j)}{\tilde{q}_j(x_j)}.$$

The corresponding PSN^2 classifier is thus given by

$$\hat{\phi}_{\mathrm{PSN}^2}(x) = \mathbb{1}\left\{\hat{r}_{\mathrm{P}}^S(x) \geq (\hat{r}_{\mathrm{P}}^S)_{(k_{\min})}\right\}. \tag{2.15}$$

We assume the domains of all p_j and q_j to be $[-1, 1]$ for all the following theoretical discussion. We will prove NP oracle inequalities for $\hat{\phi}_{\mathrm{NSN}^2}$, and those for $\hat{\phi}_{\mathrm{PSN}^2}$ can be developed similarly. Recall that by Proposition 2.4, we need an upper bound for $\|\hat{r}_N^S - r\|_\infty$. Necessarily, performance of the screening step should be studied. To this end, we assume that only a small fraction of the d features have marginal differentiating power.

Assumption 2. *There exists a signal set $\mathcal{A} \subset \{1, \cdots, d\}$ with size $|\mathcal{A}| = s \ll d$ such that $\inf_{j\in\mathcal{A}} \|F_j^0 - F_j^1\|_\infty \geq D$ for some positive constant D, and $F_j^0 = F_j^1$ for $j \notin \mathcal{A}$.*

The following proposition shows that Step 1.1 achieves exact recovery ($\widehat{\mathcal{A}}_\tau = \mathcal{A}$) with high probability for some properly chosen τ.

Proposition 2.5 (exact recovery)**.** *Let $\delta_1 \in (0,1)$. In addition to Assumptions 1 and 2, suppose $n_1 \wedge m_1 \geq 8D^{-2}\log(4d/\delta_1)$. Then for any $\tau \in [\Delta_0, D-\Delta_0]$, where $\Delta_0 = \sqrt{\frac{\log(4d/\delta_1)}{2n_1}} + \sqrt{\frac{\log(4d/\delta_1)}{2m_1}}$, the screening substep Step 1.1* (2.12) *satisfies*

$$\mathbb{P}(\widehat{\mathcal{A}}_\tau = \mathcal{A}) \geq 1 - \delta_1 \,.$$

Now we are ready to control the uniform deviation of density ratio estimate given in Step 1.2.

Assumption 3. *The marginal densities p_j, $q_j \in \mathcal{P}_\Sigma(\beta, L, [-1,1])$ for all $j = 1, \cdots, d$, and there exists $\underline{\mu} > 0$ such that $p_j, q_j \geq \underline{\mu}$ for all $j \in \mathcal{A}$. There exists some constant $\bar{C} > 0$, such that $\|r\|_\infty \leq \bar{C}$, and there is a uniform absolute upper bound for $\|p_j^{(l)}\|_\infty$ and $\|q_j^{(l)}\|_\infty$ for $j \in \mathcal{A}$ and $l \in [0, \lfloor\beta\rfloor]$. Moreover, the kernel K in the nonparametric density estimates is β-valid and L'-Lipschitz.*

Smoothness conditions (Assumption 3) and the margin assumption were used together in the classical classification literature. However, it is not entirely obvious why Assumption 3 does not render the detection condition redundant. We refer interested readers to Appendix B for more detailed discussion.

Let C_j^1 and C_j^0 be the constants C in Lemma A.6 when applied to p_j and q_j respectively. Assumption 3 ensures the existence of absolute constants $C^1 \geq \sup_{j\in\mathcal{A}} C_j^1$ and $C^0 \geq \sup_{j\in\mathcal{A}} C_j^0$.

Proposition 2.6 (uniform deviation of density ratio estimate)**.** *Under Assumptions 1 - 3, for any $\delta_1, \delta_2 \in (0,1)$, if $n_1 \wedge m_1 \geq 8D^{-2}\log(4d/\delta_1)$, $\sqrt{\frac{\log(2n_2 s/\delta_2)}{n_2 h_1}} \leq \min(1, \underline{\mu}/C^1)$, $\sqrt{\frac{\log(2m_2 s/\delta_2)}{m_2 h_0}} \leq \min(1, \underline{\mu}/C^0)$, and the screening threshold τ is specified as in Proposition 2.5, we have*

$$\mathbb{P}\left(\|\hat{r}_N^S - r\|_\infty \leq T\right) \geq 1 - \delta_1 - \delta_2 \,, \tag{2.16}$$

where $T = Be^B\|r\|_\infty$ with

$$B = s\left\{\frac{C^1\sqrt{\frac{\log(2n_2 s/\delta_2)}{n_2 h_1}}}{\underline{\mu} - C^1\sqrt{\frac{\log(2n_2 s/\delta_2)}{n_2 h_1}}} + \frac{C^0\sqrt{\frac{\log(2m_2 s/\delta_2)}{m_2 h_0}}}{\underline{\mu} - C^0\sqrt{\frac{\log(2m_2 s/\delta_2)}{m_2 h_0}}}\right\}.$$

Moreover, assume that $n_2 \wedge m_2 \geq 1/\delta_2$, $|\mathcal{A}| = s \leq (n_2 \wedge m_2)^{\frac{\beta}{2(\beta+1)}}$, and the bandwidths $h_1 = (\log n_2/n_2)^{\frac{1}{2\beta+1}}$ and $h_0 = (\log m_2/m_2)^{\frac{1}{2\beta+1}}$, then there exists an absolute constant $C_2 > 0$ such that

$$\mathbb{P}\left[\|\hat{r}_N^S - r\|_\infty \leq C_2\, s\left\{\left(\frac{\log n_2}{n_2}\right)^{\frac{\beta}{2\beta+1}} + \left(\frac{\log m_2}{m_2}\right)^{\frac{\beta}{2\beta+1}}\right\}\right] \geq 1 - \delta_1 - \delta_2 \,.$$

The condition $|\mathcal{A}| = s \leq (n_2 \wedge m_2)^{\frac{\beta}{2(\beta+1)}}$ in the above proposition ensures that the upper bound of the uniform deviation diminishes as sample sizes n_2, m_2 go to infinity. Now we are in a position to present the theorem finale of NSN^2.

Theorem 2.1 (NP Oracle Inequalities for $\hat{\phi}_{\mathrm{NSN}^2}$)**.** *In addition to Assumptions 1 - 3, assume the density ratio r satisfies the margin assumption of order $\bar{\gamma}$ at level C_α and detection condition of order $\underline{\gamma}$ at level (C_α, δ^*), both with respect to P_0. For any given $\delta_1, \delta_2, \delta_3, \delta_4 \in (0,1)$, let the NSN^2 classifier $\hat{\phi}_{NSN^2}$ be defined as in* (2.14)*, with the screening threshold τ specified as in Proposition 2.5 and kernel bandwidths $h_1 = (\log n_2/n_2)^{\frac{1}{2\beta+1}}$ and $h_0 = (\log m_2/m_2)^{\frac{1}{2\beta+1}}$, and $\hat{r}_N^S$ be such that $F_{0,\hat{r}_N^S}$ is continuous almost surely. For subsample sizes that satisfy $n_1 \wedge m_1 \geq 8D^{-2}\log(4d/\delta_1)$, $n_2 \wedge m_2 \geq \max\{\delta_2^{-1}, s^{\frac{2(\beta+1)}{\beta}}\}$, $\sqrt{\frac{\log(2n_2 s/\delta_2)}{n_2 h_1}} \leq \min(1, \underline{\mu}/C^1)$, $\sqrt{\frac{\log(2m_2 s/\delta_2)}{m_2 h_0}} \leq \min(1, \underline{\mu}/C^0)$,*
and $m_3 \geq \max\{4/(\alpha\delta_3), \delta_3^{-2}, \delta_4^{-2}, (\frac{2}{5}M_1 {\delta^*}^{\underline{\gamma}})^{-4}\}$, there exists an absolute constant $\tilde{C} > 0$ such that with probability at least $1 - \delta_1 - \delta_2 - \delta_3 - \delta_4$,

(I) $R_0(\hat{\phi}_{NSN^2}) \leq \alpha$,

(II) $R_1(\hat{\phi}_{NSN^2}) - R_1(\phi^*) \leq \tilde{C} \left\{ m_3^{-(\frac{1}{4} \wedge \frac{1+\bar{\gamma}}{4\underline{\gamma}})} + s^{1+\bar{\gamma}} \left(\frac{\log n_2}{n_2}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+1}} + s^{1+\bar{\gamma}} \left(\frac{\log m_2}{m_2}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+1}} \right\}$.

Theorem 2.1 establishes the NP oracle inequalities for $\hat{\phi}_{\mathrm{NSN}^2}$. To help understand the conditions of this theorem, recall that Assumption 1 is about sample splitting, Assumption 2 is on minimal signal strength for active feature set, Assumption 3 is on marginal densities and kernels in nonparametric estimates, and the margin assumption and detection condition describe the neighbourhood of the oracle decision boundary. Note that the subsample sizes n_1 and m_1 do not enter the upper bound for the excess type II error explicitly. Instead, we have size requirements on them so that the important features are kept with high probability $1 - \delta_1$ in the screening substep. The tolerance parameter δ_2 arises from the nonparametric estimation of densities, the parameter δ_3 is for the tolerance on violation of type I error bound, and δ_4 arises from controlling $|R_0(\hat{\phi}_{\mathrm{NSN}^2}) - R_0(\phi^*)|$.

3. Numerical investigation

In this section, we analyze two simulated examples and two real datasets to demonstrate the performance of our newly proposed NSN^2 and PSN^2 classifiers, in comparison with their corresponding non-screening counterparts (denoted as NN^2 and PN^2 respectively) as well as three popular methods under the classical framework: Gaussian Naive Bayes (nb), penalized logistic regression (pen-log), and Support Vector Machine (svm). We use R package "e1071" for nb and svm, and the R package "glmnet" for pen-log. To facilitate the presentation, we summarize the four Neyman-Pearson Naive Bayes classifiers in Table 2.

To train the classifiers in Table 2, we set $\alpha = 0.05$, $\delta_1 = 0.05$, and $\delta_3 = 0.05$ throughout this section unless specified otherwise. In Assumption 1, motivated by Proposition 2.5, we take $m_1 = \min\{10\log(4d/\delta_1), m/4\}\mathbb{1}(\text{screening})$, $n_1 = \min\{10\log(4d/\delta_1), n/2\}\mathbb{1}(\text{screening})$, $m_2 = \lfloor m/2 \rfloor - m_1$, $n_2 = n - n_1$, and $m_3 = m - \lfloor m/2 \rfloor$.

Table 2: A summary of the four Neyman-Pearson Naive Bayes classifiers.

	Screening-based	Non-screening
Non-parametric	$\hat{\phi}_{\mathrm{NSN}^2}(x) = \mathbb{1}\left\{\hat{r}^S_{\mathrm{N}}(x) \geq (\hat{r}^S_{\mathrm{N}})_{(k_{\min})}\right\}$	$\hat{\phi}_{\mathrm{NN}^2}(x) = \mathbb{1}\left\{\hat{r}_{\mathrm{N}}(x) \geq (\hat{r}_{\mathrm{N}})_{(k_{\min})}\right\}$
Parametric	$\hat{\phi}_{\mathrm{PSN}^2}(x) = \mathbb{1}\left\{\hat{r}^S_{\mathrm{P}}(x) \geq (\hat{r}^S_{\mathrm{P}})_{(k_{\min})}\right\}$	$\hat{\phi}_{\mathrm{PN}^2}(x) = \mathbb{1}\left\{\hat{r}_{\mathrm{P}}(x) \geq (\hat{r}_{\mathrm{P}})_{(k_{\min})}\right\}$

Due to the absence of information with respect to the true p and q, the theoretical screening cutoff that achieves exact recovery is not feasible in practice. We resort to an empirical permutation-based approach (Fan et al., 2011) as a substitute. Specifically, the screening substep in NSN^2 is executed as follows:

1. Combine $\mathcal{S}^0_1$ and $\mathcal{S}^1_1$ into $\{(X_i, Y_i)\}_{i=1}^{m_1+n_1}$, where $X_i \in \mathcal{S}^0_1 \cup \mathcal{S}^1_1$, and Y_i is X_i's class label.

2. Calculate the marginal D-statistic for each feature:
$$D_j = \|\hat{F}^0_j - \hat{F}^1_j\|_\infty, \quad j = 1, 2, \cdots, d\,,$$
where $\hat{F}^0_j(x) = \sum_{i:Y_i=0} \mathbb{1}(X_{i,j} \leq x_j)$ and $\hat{F}^1_j(x) = \sum_{i:Y_i=1} \mathbb{1}(X_{i,j} \leq x_j)$.

3. Let $\pi = \{\pi(1), \cdots, \pi(m_1 + n_1)\}$ be a random permutation of $\{1, \cdots, (m_1 + n_1)\}$. For $j = 1, \cdots, d$, compute $D^{\mathrm{null}}_j = \|\hat{F}^{0,\mathrm{null}}_j - \hat{F}^{1,\mathrm{null}}_j\|_\infty$, where $\hat{F}^{0,\mathrm{null}}_j(x_j) = \sum_{i:Y_{\pi(i)}=0} \mathbb{1}(X_{i,j} \leq x_j)$, $\hat{F}^{1,\mathrm{null}}_j(x_j) = \sum_{i:Y_{\pi(i)}=1} \mathbb{1}(X_{i,j} \leq x_j)$.

4. For some pre-specified $Q \in [0, 1]$, let $\omega(Q)$ be the Q-th quantile of $\{D^{\mathrm{null}}_j : j = 1, \cdots, d\}$ and select $\widehat{\mathcal{A}} = \{j : D_j \geq \omega(Q)\}$. Here, Q is a tuning parameter that keeps the percentage of noise features that pass the screening around $1 - Q$.

The same permutation idea is applied to the screening substep of PSN^2. Q is set at 0.95 throughout this section.

3.1 Simulation

Samples in both simulated examples are generated from the model
$$p(x) = \prod_{j=1}^{d} p_j(x_j), \quad q(x) = \prod_{j=1}^{d} q_j(x_j)$$
at 3 different dimensions: $d \in \{10, 100, 1000\}$. Sparsity for $d = 100$ and 1000 is imposed by setting $p_j = q_j$ for all $j > 10$. Seven different training sample sizes: $m = n \in \{$200, 400, 800, 1600, 3200, 6400, 12800$\}$ are considered. The number of replications for each scenario is 1000. Test errors are estimated using the average of 1000 independent observations from each class for each replication.

3.1.1 Example 1: normals with different means

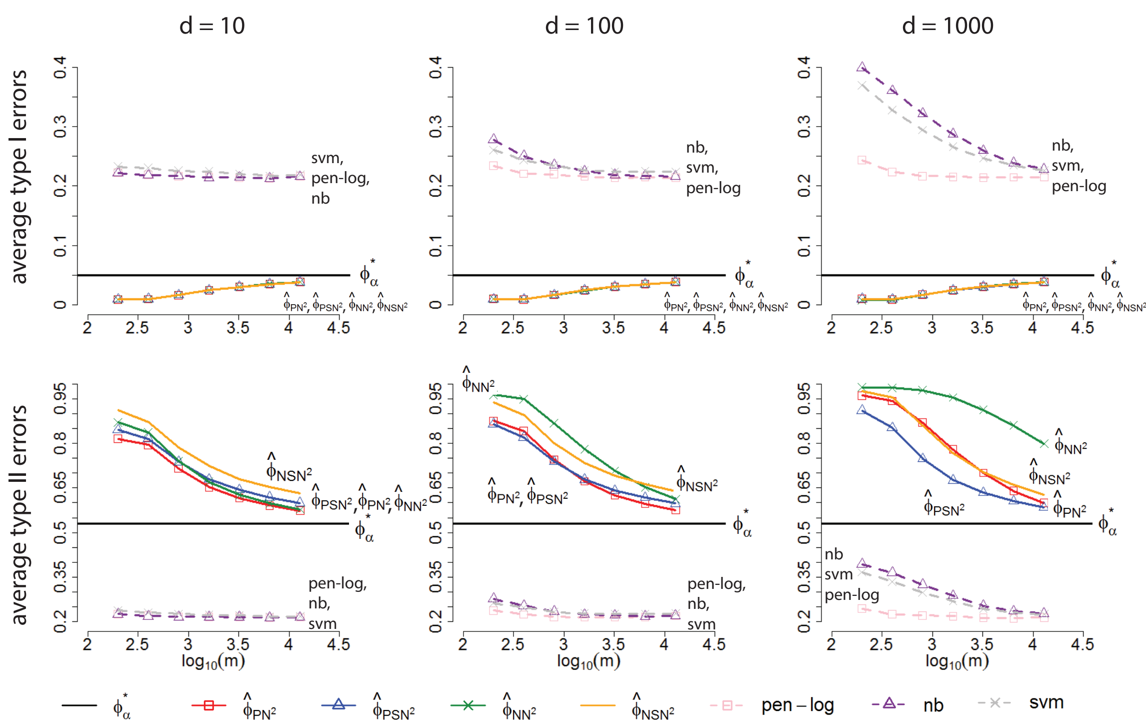
Assume the two-class conditional densities $p \sim \mathcal{N}(0.5(1'_{10}, 0'_{d-10})', I_d)$ and $q \sim \mathcal{N}(0_d, I_d)$ where I_d is the identity matrix. At significance level $\alpha = 0.05$, the oracle type I/II risks are $R_0(\phi^*_\alpha) = 0.05$ and $R_1(\phi^*_\alpha) = 0.53$ respectively.

We first evaluate the screening performance of PSN^2 and NSN^2 with results presented in Table 3. Both t-statistic (in PSN^2) and D-statistic (in NSN^2) are able to pick up most of the true signals while keeping the false positive rates at around $1 - Q$.

Table 3: Average screening performance summarized over 1000 independent replications at sample sizes $m = n = 400$ and $Q = 0.95$ with standard errors in parentheses.

	# of selected features		# of missed signals		# of false positive	
d	t-stat	D-stat	t-stat	D-stat	t-stat	D-stat
10	9.11 (1.14)	8.11 (1.63)	0.89 (1.14)	1.89 (1.63)	0 (0)	0 (0)
100	14.64 (3.46)	12.43 (3.38)	0.78 (0.90)	2.00 (1.39)	5.43 (3.17)	4.43 (2.77)
1000	59.99 (9.77)	58.82 (9.87)	0.48 (0.66)	1.14 (1.05)	50.47 (9.71)	49.96 (9.78)

Figure 1: Average errors of $\hat{\phi}$'s over 1000 independent replications for each combination of (d, m, n).



We then move on to evaluate the trend of type I and type II errors as the sample size increases in Figure 1. All the Neyman-Pearson based classifiers have type I error

approaching α from below as sample size increases and they have similar type I errors at each sample size. However, nb, pen-log and svm all lead to a type I error larger than α.

By enlarging the second row of Figure 1, one would observe the differences in type II errors among PN^2, PSN^2, NN^2, NSN^2. In the case of $d = 10$ when all features are signals, PN^2 performs the best throughout all sample sizes since it assumes the correct model without the unnecessary screening substep. When sample size is small, PSN^2 outperforms NN^2, but NN^2 gradually catches up on larger samples. In the case of $d = 100$, screening helps PSN^2 to take the lead at low sample sizes. The advantage of screening fades off as the sample size increases. In the case of $d = 1000$, PSN^2 dominates all other three classifiers throughout the sample size range we investigate.

Overall, the advantage of PSN^2 over NSN^2, and PN^2 over NN^2 are uniform across all dimensions and sample sizes. This is consistent with the intuition that when the data are from a two-class Gaussian model, the parametric methods lead to more efficient estimators than nonparametric counterparts.

3.1.2 Example 2: normal vs. mixture normal

Normality assumption is violated in the second example. Assume $p \sim 0.5\mathcal{N}(a, \Sigma) + 0.5\mathcal{N}(-a, \Sigma)$ and $q \sim \mathcal{N}(0_d, I_d)$, where $a = (\frac{3}{\sqrt{10}} 1'_{10}, 0'_{d-10})'$, $\Sigma = \begin{pmatrix} 10^{-1} I_{10} & 0 \\ 0 & I_{d-10} \end{pmatrix}$. At significance level $\alpha = 0.05$, the oracle type I/II risks are $R_0(\phi^*_\alpha) = 0.05$ and $R_1(\phi^*_\alpha) = 0.027$ respectively.

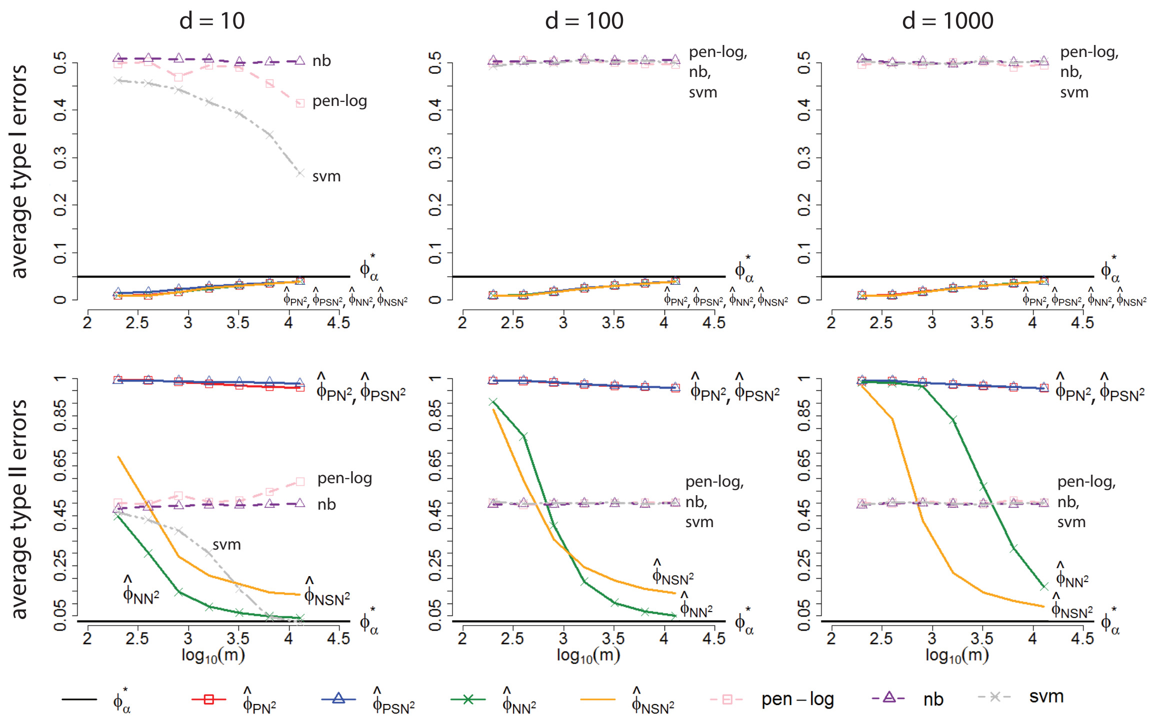
The performance of the screening substep of PSN^2 and NSN^2 is shown in Table 4. While both screening methods keep the false positive rates at around $1 - Q$, the parametric screening method (PSN^2) with t-statistic misses almost all signals. This is not surprising since t-statistics rank features by differences in means and the two groups have exactly the same marginal mean and variance across all dimensions.

Table 4: Average screening performance summarized over 1000 independent replications at sample sizes $m = n = 400$ and $Q = 0.95$ with standard errors in parentheses.

	# of selected features		# of missed signals		# of false positive	
d	t-stat	D-stat	t-stat	D-stat	t-stat	D-stat
10	1.76 (1.53)	8.13 (1.83)	8.24 (1.53)	1.87 (1.83)	0 (0)	0 (0)
100	5.93 (3.44)	11.96 (3.57)	9.38 (0.80)	2.34 (1.59)	5.31 (3.17)	4.29 (2.68)
1000	50.69 (9.60)	58.78 (9.87)	9.50 (0.69)	1.26 (1.04)	50.19 (9.51)	50.04 (9.62)

Figure 2 presents the average error rates. The same reason that causes the above fiasco of t-statistic screening reduces PSN^2 and PN^2 to nothing more than, if not less than, two unfair random coins with probability 0.05 of landing 1, while the behaviors of nb and pen-log bear more resemblance to that of fair random coins. This fundamental difference is due to that the classical framework aims to minimize the overall risk, and therefore tends to distribute errors evenly when the sample size for the two classes are about the same. The NSN^2 and NN^2 based on nonparametric assumptions, on the other hand, perform very well on non-normal data. Their difference in type II error performances are similar as in Example 1.

Figure 2: Average error rates of $\hat{\phi}$'s over 1000 independent replications for each combination of (d, m, n). Error rates are computed as the average of 1000 independent testing data points from each class in each replication, and then average over replications.



From the two simulation examples, it is clear that the screening-based NSN^2 and PSN^2 exhibit advantages over their non-screening counterparts under high-dimensional settings. When the normality assumption is violated, and the sample sizes are reasonably large for efficient kernel estimates, NSN^2 prevails over PSN^2 . As a rule of thumb, for high-dimensional classification problems that emphasize type I error control, we recommend NSN^2 if the sample size is relatively large and PSN^2 otherwise.

3.2 Real data analysis

In addition to the neuroblastoma dataset analyzed in the introduction, we now demonstrate the performance of PSN^2 and NSN^2 for targeted asymmetric error control on two additional real datasets.

3.2.1 p53 mutants dataset

The p53 mutants dataset (Danziger et al., 2006) contains $d = 5407$ attributes extracted from biophysical experiments for 16772 mutant p53 proteins, among which 143 are determined as "active" and the rest as "inactive" via in vivo assays.

All 143 active samples and the first 1500 inactive samples are included in our analysis. We treat the active class as class 0 and aimed to control the error of missing an active under $\alpha = 0.05$. This dataset is split into a training set with 100 observations from the active class and 1000 observations from the inactive class, and a testing set with the remaining observations. PSN^2 is used as the representative of our proposed methods, as the class 0 sample size is small for nonparametric methods. The average type I and type II errors over 1000 random splits are shown in Table 5. Compared with pen-log, nb and svm, PSN^2 performs much better in controlling the type I error.

Table 5: Average errors over 1000 random splits with standard errors in parentheses. $\alpha = 0.05$, $\delta_1 = 0.05$, $Q = 0.95$, and $\delta_3 = 0.1$.

	PSN^2	pen-log	nb	svm
type I	$\underline{.019\ (.028)}$	.162 (.060)	.056 (.034)	.484 (.222)
type II	.461 (.291)	.010 (.004)	.458 (.033)	.004 (.003)

3.2.2 Email spam dataset

Now, we consider an e-mail spam dataset available at `https://archive.ics.uci.edu/ml/datasets/Spambase`, which contains 4601 observations with 57 features, among which 2788 are class 0 (non-spam) and 1813 are class 1 (spam). We first standardize each feature and add 5000 synthetic features consisting of independent $\mathcal{N}(0, 1)$ variables to make the problem more challenging. The augmented data has $n = 4601$ observations with $d = 5057$ features. This augmented dataset is split into a training set with 1000 observations from each class and a testing set with the remaining observations. We use NSN^2 since the sample size is relatively large. The average type I and type II errors over 1000 random splits are shown in Table 6.

To evaluate the flexibility of NSN^2 in terms of prioritized error control, we also report the performance when the priority is switched to control the type II error below $\alpha = 0.05$. The results in Table 6 demonstrate that NSN^2 is able to control either type I or type II error depending on the specific need of the practitioner.

Table 6: Average errors over 1000 random splits with standard errors in parentheses. $\alpha = 0.05$, $\delta_1 = 0.05$, $Q = 0.95$, and $\delta_3 = 0.05$. The suffix after NSN^2 indicates the type of error it targets to control under α.

	NSN^2-R_0	NSN^2-R_1	pen-log	nb	svm
type I	$\underline{.019\ (.007)}$	.488 (.078)	.064 (.007)	.444 (.018)	.203 (.013)
type II	.439 (.057)	$\underline{.020\ (.009)}$	.133 (.015)	.054 (.008)	.235 (.017)

4. Discussion

The Neyman-Pearson classification framework is an important and interesting paradigm to explore beyond the Naive Bayes models considered in this work. For example, we can relax the independence assumption on PSN^2, and consider a general covariance matrix. Also, we can consider NP-type classifiers with decision boundaries involving feature interactions. It is also worthwhile to study the non-probabilistic approaches under high-dimensional NP paradigm. Methods of potential interest include the k nearest neighbor (Weiss et al., 2010) and the centroid based classifiers (Tibshirani et al., 2002; Hall et al., 2010). However, the NP oracle inequalities are likely to be replaced by a new theoretical formulation for these methods.

A benefit of the present approach is that, for any given estimator $\hat{r}$, we have a uniform method to determine the proper threshold level in the plug-in classifiers. However, it would be interesting to develop new ways to estimate the threshold level C_α that is adaptive to the particular method used to approximate the density ratio r.

Appendix A. Technical Lemmas and Proofs

Let $\mathrm{Bin.cdf}_{n,\,p}(\cdot)$ denote the CDF of $\mathrm{Bin}(n,p)$, and $\mathrm{Beta.cdf}_{a,\,b}(\cdot)$ denote the CDF of $\mathrm{Beta}(a,b)$. The following lemma proves a duality between the beta and binomial distributions.

Lemma A.1 (Beta-binomial duality)**.** *For any $p \in [0,1]$ and $k \in \{1,\dots,n\}$, it holds that*

$$1 - \mathit{Bin.cdf}_{n,\,p}(k-1) \;=\; \mathit{Beta.cdf}_{k,\,n+1-k}\,(p)\,.$$

Proof of Lemma A.1. Let $U_1,\dots,U_n$ be n i.i.d. Uniform$[0,1]$. For any $p \in [0,1]$, let $N_p = \sum_{i=1}^{n} \mathbb{1}\{U_i \le p\}$ denote the number of U_i's that are less or equal to p. Given

$$\mathbb{P}(\mathbb{1}\{U_i \le p\} = 1) \;=\; \mathbb{P}(U_i \le p) \;=\; p\,, \quad \mathbb{1}\{U_i \le p\} \;\sim\; \mathrm{Bern}(p) \quad \forall i\,,$$

we have $N_p \sim \mathrm{Bin}(n,p)$, and therefore

$$\mathbb{P}\left(N_p \ge k\right) \;=\; 1 - \mathbb{P}\left(N_p \le k-1\right) \;=\; 1 - \mathrm{Bin.cdf}_{n,p}(k-1)\,. \tag{A.1}$$

On the other hand, let $U_{(k)}$ denote the k-th order statistic of $\{U_i\}_{i=1}^n$. It follows from the definition of order statistics that

$$\{N_p \ge k\} \;=\; \{\text{at least } k \text{ of } U_1,\dots,U_n \text{ are less or equal to } p\} \;=\; \left\{U_{(k)} \le p\right\}\,. \tag{A.2}$$

Combining (A.1) with (A.2) yields

$$1 - \mathrm{Bin.cdf}_{n,p}(k-1) \;=\; \mathbb{P}\left(N_p \ge k\right) \;=\; \mathbb{P}\left(U_{(k)} \le p\right) \;=\; \mathrm{Beta.cdf}_{k,\,n+1-k}\,(p)\,,$$

where the last equality follows from $U_{(k)} \sim \mathrm{Beta}(k, n+1-k)$ $(k = 1,\dots,n)$ as a direct implication of Rényi's representation. This completes the proof. $\square$

Lemma A.2. *Let Z be a random variable from* CDF *F. We have*

$$P_F\left\{F(Z) < \delta\right\} \;\le\; \delta\,, \quad P_F\left\{F(Z) > \delta\right\} \;\ge\; 1-\delta \quad \forall \delta \in [0,1]\,. \tag{A.3}$$

For continuous F, the inequality becomes equality as

$$P_F\left\{F(Z) < \delta\right\} \;=\; \delta\,, \quad P_F\left\{F(Z) > \delta\right\} \;=\; 1-\delta \quad \forall \delta \in [0,1]\,. \tag{A.4}$$

.

Proof. Let $t_1 = \min\{t : F(t) \geq \delta\}$. Given the right continuity of F, it can be easily proved by contradiction that i) $F(t_1-) = F(t_1) = \delta$ if F is continuous at t_1, and ii) $F(t_1-) < \delta \leq F(t_1)$ if F is discontinuous at t_1. Thus,

$$P_F\{F(Z) < \delta\} \;=\; P_F(Z < t_1) \;=\; F(t_1-) \;\leq\; \delta\,.$$

Likewise, let $t_2 = \inf\{t : F(t) > \delta\}$. We have i) $F(t_2-) = F(t_2) = \delta$ if F is continuous at t_2, and ii) $F(t_2-) < \delta \leq F(t_2)$ if F is discontinuous at t_2. As a result,

$$P_F\{F(Z) > \delta\} \;=\; P_F\left\{Z \geq t_2\right\} \;=\; 1 - P_F\left\{Z < t_2\right\} \;\geq\; 1-\delta\,.$$

This completes the proof. □

Lemma A.3. *Let $\mathcal{S} = \{Z_i\}_{i=1}^n$ be a set n i.i.d. random variables from distribution F, and let $Z_{(k)}$ denote its k-th order statistic ($k = 1,\ldots,n$). For any $\delta \in (0,1)$, the probability of a new, independent realization Z from F to be greater than $Z_{(k)}$ satisfies*

$$\mathbb{P}\left\{P_F\left(Z > Z_{(k)} \mid \mathcal{S}\right) > \delta\right\} \;\leq\; 1 - \mathit{Bin.cdf}_{n,1-\delta}(k-1)\,, \tag{A.5}$$

$$\mathbb{P}\left\{P_F\left(Z > Z_{(k)} \mid \mathcal{S}\right) < \delta\right\} \;\geq\; 1 - \mathit{Bin.cdf}_{n,\delta}(n-k) \;=\; \mathit{Bin.cdf}_{n,1-\delta}(k-1)\,. \tag{A.6}$$

The inequalities become equalities if F is continuous.

Proof of Lemma A.3. Rewrite the left-hand side of (A.5) as

$$\begin{aligned} &\mathbb{P}\left\{P_F\left(Z > Z_{(k)} \mid \mathcal{S}\right) > \delta\right\} \;=\; \mathbb{P}\left\{1 - P_F\left(Z \leq Z_{(k)} \mid \mathcal{S}\right) > \delta\right\} \\ =\;& \mathbb{P}\left\{1 - F(Z_{(k)}) > \delta\right\} \;=\; \mathbb{P}\left\{F(Z_{(k)}) < 1-\delta\right\}. \end{aligned} \tag{A.7}$$

To bound the probability of $\{F(Z_{(k)}) < 1-\delta\}$, let $N_{1-\delta} \;=\; \sum_{i=1}^n \mathbb{I}_{\{F(Z_i)<1-\delta\}}$ denote the number of $F(Z_i)$'s that are less than $1-\delta$. It follows from $F(Z_{(1)}) \leq F(Z_{(2)}) \leq \ldots \leq F(Z_{(n)})$ that

$$\begin{aligned} &\left\{F(Z_{(k)}) < 1-\delta\right\} \;=\; \left\{F(Z_{(i)}) < 1-\delta\,,\; i = 1,\ldots,k\right\} \;=\; \left\{N_{1-\delta} \geq k\right\}\,, \\ &\mathbb{P}\left\{F(Z_{(k)}) < 1-\delta\right\} \;=\; \mathbb{P}\left(N_{1-\delta} \geq k\right)\,. \end{aligned} \tag{A.8}$$

Let $\tau = P_F\left\{F(Z_1) < 1-\delta\right\}$ denote the success probability of $N_{1-\delta}$ as a binomial. It follows from (A.3) that $\tau \leq 1-\delta$. Given $\mathrm{Bin.cdf}_{n,p}(k-1)$ being decreasing in p for any fixed n and k, we have

$$\mathbb{P}\left(N_{1-\delta} \geq k\right) \;=\; 1 - \mathrm{Bin.cdf}_{n,\tau}(k-1) \;\leq 1 - \mathrm{Bin.cdf}_{n,\,1-\delta}(k-1) \tag{A.9}$$

as a result of The equalities hold for continuous F. Connecting (A.7), (A.8), and (A.9) together yields

$$\begin{aligned} \mathbb{P}\left\{P_F\left(Z > Z_{(k)} \mid \mathcal{S}\right) > \delta\right\} \;&=\; \mathbb{P}\left\{F(Z_{(k)}) < 1-\delta\right\} \;=\; \mathbb{P}\left(N_{1-\delta} \geq k\right) \\ &\leq\; 1 - \mathrm{Bin.cdf}_{n,\,1-\delta}(k-1)\,. \end{aligned}$$

Likewise, let $M_{1-\delta} = \sum_{i=1}^{n} \mathbb{I}_{\{F(Z_i)>1-\delta\}}$ be a binomial random variable with size n and success rate $\tau' = P_F\{F(Z_i) > 1-\delta\} \geq \delta$ that represents the number of $F(Z_i)$'s that are greater than $1-\delta$. The left-hand side of (A.6) can be rewritten as

$$\begin{aligned}
& \mathbb{P}\left\{P_F\left(Z > Z_{(k)} \mid \mathcal{S}\right) < \delta\right\} = \mathbb{P}\left\{F(Z_{(k)}) > 1-\delta\right\} \\
= \; & \mathbb{P}\left\{F(Z_{(i)}) > 1-\delta\,,\; i = k, \ldots, n\right\} = \mathbb{P}\left\{M_{1-\delta} \geq n+1-k\right\} \\
= \; & 1 - \mathbb{P}\left\{M_{1-\delta} \leq n-k\right\} = 1 - \text{Bin.cdf}_{n,\tau'}(n-k) \\
\geq \; & 1 - \text{Bin.cdf}_{n,\,\delta}(n-k). \qquad\qquad (\text{A.10})
\end{aligned}$$

This completes the proof. □

Proof of Proposition 2.1. Letting $Z_i = \hat{r}_i$, $n = m_3$ in Lemma (A.3) yields

$$\mathbb{P}\{R_0(\hat{\phi}_k) > \delta\} \quad \leq \quad 1 - \text{Bin.cdf}_{m_3,\,1-\delta}(k-1)\,.$$

This, together with Lemma A.1, completes the proof. □

Lemma A.4. *For random variable $Z \sim$ Beta(a, b), and any $\epsilon > 0$, we have*

$$\mathbb{P}\{Z > (1+\epsilon)\mathbb{E}Z\} < \mathbb{P}(|Z - \mathbb{E}Z| > \epsilon\mathbb{E}Z) < \frac{b\epsilon^{-2}}{(a+b+1)a}. \qquad (\text{A.11})$$

Proof of Lemma A.4. By Chebyshev inequality,

$$\mathbb{P}(|Z - \mathbb{E}Z| > \epsilon\mathbb{E}Z) \leq \frac{\text{var}(Z)}{(\epsilon\mathbb{E}Z)^2} = \frac{ab}{(a+b)^2(a+b+1)}\left(\frac{\epsilon a}{a+b}\right)^{-2} = \frac{b\epsilon^{-2}}{(a+b+1)a}.$$

□

Proof of Proposition 2.2. Let B be a realization from Beta(k, m_3+1-k). It follows from Proposition 2.1 that

$$\begin{aligned}
\mathbb{P}\{R_0(\hat{\phi}_k) > g(\delta_3, m_3, k)\} \quad & \leq \quad \text{Beta.cdf}_{k,m_3+1-k}\left\{1 - g(\delta_3, m_3, k)\right\} \\
& = \quad \mathbb{P}\{B \leq 1 - g(\delta_3, m_3, k)\} \; = \; \mathbb{P}\{1 - B \geq g(\delta_3, m_3, k)\}
\end{aligned}$$

for any $k \in \{1, \cdots, m_3\}$ and $\hat{r}$, with $1 - B \sim$ Beta(m_3+1-k, k). Letting $a = m_3+1-k$, $b = k$, and $\epsilon = k^{1/2}\{\delta_3(m_3+2)(m_3+1-k)\}^{-1/2}$ in Lemma A.4 yields

$$\mathbb{P}\{R_0(\hat{\phi}_k) > g(\delta_3, m_3, k)\} \;\leq\; \delta_3\,,$$

where

$$g(\delta_3, m_3, k) = (1+\epsilon)\left(\frac{m_3+1-k}{m_3+1}\right) = \frac{m_3+1-k}{m_3+1} + \sqrt{\frac{k(m_3+1-k)}{\delta_3(m_3+2)(m_3+1)^2}}\,.$$

This completes the proof.

□

Proof of Proposition 2.3. By some basic algebra we have

$$A_{\alpha,\delta_3}(m_3) - (1-\alpha) = \frac{-1+2\alpha+\sqrt{1+4\delta_3(1-\alpha)\alpha(m_3+2)}}{2\{\delta_3(m_3+2)+1\}} > 0\,,$$
$$A_{\alpha,\delta_3}(m_3) - 1 = \frac{-1-2\delta_3(m_3+2)\alpha+\sqrt{1+4\delta_3(1-\alpha)\alpha(m_3+2)}}{2\{\delta_3(m_3+2)+1\}} < 0\,,$$

and

$$g(\delta_3,m_3,k) = \frac{m_3+1-k}{m_3+1} + \sqrt{\frac{k(m_3+1-k)}{\delta_3(m_3+2)(m_3+1)^2}} \le \alpha$$
$$\Leftrightarrow \quad \begin{cases} k-(1-\alpha)(m_3+1) \ge 0, \\ \\ \{\delta_3(m_3+2)+1\}\left(\frac{k}{m_3+1}\right)^2 - \{1+2\delta_3(m_3+2)(1-\alpha)\}\left(\frac{k}{m_3+1}\right) \\ +\delta_3(m_3+2)(1-\alpha)^2 \ge 0 \end{cases}$$
$$\Leftrightarrow \quad k \ge (m_3+1)\max\{1-\alpha, A_{\alpha,\delta_3}(m_3)\}$$
$$\Leftrightarrow \quad k \ge (m_3+1)A_{\alpha,\delta_3}(m_3)\,.$$

Thus,

$$\begin{aligned} k_{\min}(\alpha,\delta_3,m_3) &= \lceil (m_3+1)A_{\alpha,\delta_3}(m_3)\rceil \\ &\in \left[(m_3+1)A_{\alpha,\delta_3}(m_3), (m_3+1)A_{\alpha,\delta_3}(m_3)+1\right]\,. \end{aligned}$$

Since $A_{\alpha,\delta_3}(m_3) \to 1-\alpha$, as $m_3 \to \infty$, it follows from sandwich rule that

$$\lim_{m_3\to\infty} \frac{k_{\min}(\alpha,\delta_3,m_3)}{m_3} \;=\; \lim_{m_3\to\infty} A_{\alpha,\delta_3}(m_3) \;=\; 1-\alpha\,.$$

We have $k_{\min}(\alpha,\delta_3,m_3) \in \mathcal{K}(\alpha,\delta_3,m_3)$ ($\Leftrightarrow k_{\min}(\alpha,\delta_3,m_3) \le m_3$) as long as

$$(m_3+1)A_{\alpha,\delta_3}(m_3)+1 \;\le\; m_3 \quad \Leftrightarrow \quad (1-\alpha\;\le)\;\; A_{\alpha,\delta_3}(m_3) \;\le\; \frac{m_3-1}{m_3+1}\,. \tag{A.12}$$

For any $\Delta \in (0,\alpha)$, a sufficient condition for (A.12) is

$$\frac{m_3-1}{m_3+1} \ge 1-\Delta, \quad A_{\alpha,\delta_3}(m_3) \le 1-\Delta\,,$$

which can be further simplified as

$$m_3 \ge \frac{2}{\Delta} - 1, \quad m_3 \ge x^* - 2\,,$$

where

$$x^* \;=\; \frac{-2\Delta^2-\alpha^2+2\alpha\Delta+\Delta+(1-2\alpha)\Delta+\alpha^2}{2(\alpha-\Delta)^2\delta_3} = \frac{\Delta(1-\Delta)}{(\alpha-\Delta)^2\delta_3}$$

is the positive root of the quadratic equation

$$(\alpha-\Delta)^2\delta_3^2x^2+\delta_3\left(2\Delta^2+\alpha^2-2\alpha\Delta-\Delta\right)x-\Delta(1-\Delta)=0\,.$$

Thus, a sufficient condition for (A.12) is

$$m_3\geq\max\left\{\frac{\Delta(1-\Delta)}{(\alpha-\Delta)^2\delta_3}-2,\;\frac{2}{\Delta}-1\right\}\,.$$

Setting $\Delta=\alpha/2$ yields

$$\max\left\{\frac{\Delta(1-\Delta)}{(\alpha-\Delta)^2}\cdot\frac{1}{\delta_3}-2,\frac{2}{\Delta}-1\right\}=\max\left\{\frac{2-\alpha}{\alpha\delta_3}-2,\frac{4}{\alpha}-1\right\}\leq\frac{4}{\alpha\delta_3}\,.$$

Therefore, $m_3\geq 4/(\alpha\delta_3)$ guarantees (A.12) and $k_{\min}(\alpha,\delta_3,m_3)\in\mathcal{K}(\alpha,\delta_3,m_3)$. This completes the proof. □

Proof of Lemma 2.1. Introduce shorthand notation let $A=A_{\alpha,\delta_3}(m_3)$ (defined in Proposition 2.3) and $\alpha_1=(m_3{+}1{-}k_{\min})/(m_3{+}1)$ for simplicity of exposition. For any $B_1,B_2\in\mathbb{R}^+$, we have

$$\{|R_0(\hat\phi)-\alpha|>B_1+B_2\}\;\subset\;\{|R_0(\hat\phi)-\alpha_1|>B_1\}\cup\{|\alpha_1-\alpha|>B_2\}\,,$$

and thus

$$\begin{aligned}&\mathbb{P}\{|R_0(\hat\phi)-\alpha|>B_1+B_2\mid\hat r\}\\&\leq\;\mathbb{P}\{|R_0(\hat\phi)-\alpha_1|>B_1\mid\hat r\}+\mathbb{P}\left(|\alpha_1-\alpha|>B_2\mid\hat r\right)\\&\leq\;\frac{k_{\min}(m_3+1-k_{\min})}{(m_3+2)(m_3+1)^2}B_1^{-2}+\mathbb{1}\left\{|\alpha_1-\alpha|>B_2\right\},\end{aligned}\tag{A.13}$$

where the last inequality follows from applying Lemma A.4 to $R_0(\hat\phi)$ which follows Beta$(m_3+1-k_{\min},k_{\min})$ for $m_3\geq 4/(\alpha\delta_3)$ and continuous $F_{\hat r}$ due to Lemma A.3. It follows from Proposition 2.3 that

$$|\alpha-\alpha_1|\;\leq\;A-(1-\alpha)+\frac{1}{m_3+1}\,.\tag{A.14}$$

Letting $B_1=\sqrt{\frac{k_{\min}(m_3+1-k_{\min})}{(m_3+2)(m_3+1)^2\delta_4}}$ and $B_2=A-(1-\alpha)+\frac{1}{m_3+1}$ in (A.13) yields

$$\begin{aligned}&\mathbb{P}\{|R_0(\hat\phi)-\alpha|>\xi_{\alpha,\delta_3,m_3}(\delta_4)\mid\hat r\}\\&\leq\;\delta_4+\mathbb{1}\{|\alpha_1-\alpha|>A-(1-\alpha)+\frac{1}{m_3+1}\}\\&=\;\delta_4\end{aligned}$$

for any arbitrary $\hat r$. This, together with the independence between $\mathcal{S}_3^0$ and $\hat r$ (as a function of $(\mathcal{S}_1^0,\mathcal{S}_1^1,\mathcal{S}_2^0,\mathcal{S}_2^1)$) yields

$$\mathbb{P}\{|R_0(\hat\phi)-\alpha|>\xi_{\alpha,\delta_3,m_3}(\delta_4)\}\;\leq\;\delta_4\,.$$

To establish an upper bound for $\xi_{\alpha,\delta_3,m_3}(\delta_4)$, note that

$$
\begin{aligned}
&\xi_{\alpha,\delta_3,m_3}(\delta_4)\\
&= \sqrt{\frac{k_{\min}(m_3+1-k_{\min})}{(m_3+2)(m_3+1)^2\delta_4}} + \frac{-1+2\alpha+\sqrt{1+4\delta_3(1-\alpha)\alpha(m_3+2)}}{2\{\delta_3(m_3+2)+1\}} + \frac{1}{m_3+1}\\
&\le \sqrt{\frac{(m_3+1)^2/4}{(m_3+2)(m_3+1)^2\delta_4}} + \frac{1}{2\{\delta_3(m_3+2)+1\}} + \frac{\sqrt{1+\delta_3(m_3+2)}}{2\{\delta_3(m_3+2)+1\}} + \frac{1}{m_3+1}\\
&< \frac{1}{2\sqrt{m_3\delta_4}} + \frac{1}{2m_3\delta_3} + \frac{1}{2\sqrt{m_3\delta_3}} + \frac{1}{m_3}\,.
\end{aligned}
$$

When $m_3 \ge \max(\delta_3^{-2}, \delta_4^{-2})$, we have

$$
\begin{aligned}
\xi_{\alpha,\delta_3,m_3}(\delta_4) &< \frac{1}{2m_3^{1/4}} + \frac{1}{2m_3^{1/2}} + \frac{1}{2m_3^{1/4}} + \frac{1}{m_3}\\
&= \frac{1}{m_3^{1/4}}\left(1 + \frac{1}{2m_3^{1/4}} + \frac{1}{m_3^{3/4}}\right) < \frac{5/2}{m_3^{1/4}} = \left(\frac{2}{5}m_3^{1/4}\right)^{-1}.
\end{aligned}
$$

This completes the proof. $\square$

Proof of Proposition 2.4. Let $G^* = \{r < C_\alpha\}$ and $\widehat{G} = \{\hat{r} < \widehat{C}_\alpha\}$, the excess type II error can be decomposed as:

$$
\begin{aligned}
&P_1(\widehat{G}) - P_1(G^*)\\
&= \int_{\widehat{G}} dP_1 - \int_{G^*} dP_1 = \int_{\widehat{G}} \frac{p}{q}\, dP_0 - \int_{G^*} \frac{p}{q}\, dP_0\\
&= \int_{\widehat{G}} (r - C_\alpha) dP_0 + C_\alpha P_0(\widehat{G}) - \int_{G^*} (r - C_\alpha) dP_0 - C_\alpha P_0(G^*)\\
&= \int_{\widehat{G}\setminus G^*} (r - C_\alpha) dP_0 - \int_{G^*\setminus\widehat{G}} (r - C_\alpha) dP_0 + C_\alpha\{P_0(\widehat{G}) - P_0(G^*)\}\\
&= \int_{\widehat{G}\setminus G^*} |r - C_\alpha|\, dP_0 + \int_{G^*\setminus\widehat{G}} |r - C_\alpha|\, dP_0 + C_\alpha\{R_0(\phi^*) - R_0(\hat{\phi})\}\,. \qquad \text{(A.15)}
\end{aligned}
$$

It follows from Lemma 2.1 that when $m_3 \ge \max\{\frac{4}{\alpha\delta_3}, \delta_3^{-2}, \delta_4^{-2}, (\frac{2}{5}M_1{\delta^*}^{\gamma})^{-4}\}$,

$$
\xi_{\alpha,\delta_3,m_3}(\delta_4) \le \frac{5}{2}m_3^{-1/4} \le M_1(\delta^*)^{\gamma}, \quad \left\{\frac{\xi_{\alpha,\delta_3,m_3}(\delta_4)}{M_1}\right\}^{1/\gamma} \le \delta^*\,.
$$

Introduce shorthand notations $\Delta R_0 = |R_0(\phi^*) - R_0(\hat{\phi})|$, $\mathcal{E}_0 = \{\Delta R_0 < \xi_{\alpha,\delta_3,m_3}(\delta_4)\}$, and $T = \|\hat{r} - r\|_\infty$. On the event $\mathcal{E}_0$,

$$
\left(\frac{\Delta R_0}{M_1}\right)^{1/\gamma} \le \left\{\frac{\xi_{\alpha,\delta_3,m_3}(\delta_4)}{M_1}\right\}^{1/\gamma} \le \delta^*\,.
$$

By the detection condition, we have

$$
\Delta R_0 \le P_0\{C_\alpha < r(X) < C_\alpha + (\Delta R_0/M_1)^{1/\gamma}\}\,.
$$

Note that

$$
\begin{aligned}
P_0\{r(X) \geq C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}}\} = \; & R_0(\phi^*) - P_0\{C_\alpha < r(X) < C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}}\} \\
\leq \; & R_0(\phi^*) - \Delta R_0 \\
\leq \; & R_0(\hat{\phi}) \; = \; P_0\{\hat{r}(X) > \widehat{C}_\alpha\} \\
\leq \; & P_0\{r(X) + T \geq \widehat{C}_\alpha\} \; = \; P_0\{r(X) \geq \widehat{C}_\alpha - T\} \, .
\end{aligned}
$$

Thus, we have $\widehat{C}_\alpha \leq C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}} + T$, and

$$
\begin{aligned}
\widehat{G}\backslash G^* = \; & \{r \geq C_\alpha, \hat{r} < \widehat{C}_\alpha\} = \{r \geq C_\alpha, \hat{r} < C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}} + T\} \cap \{\hat{r} < \widehat{C}_\alpha\} \\
& = \{C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}} + 2T \geq r \geq C_\alpha, \hat{r} < C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}} + T\} \cap \{\hat{r} < \widehat{C}_\alpha\} \\
& \subset \{C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}} + 2T \geq r \geq C_\alpha\} \, .
\end{aligned}
$$

Therefore, the margin assumption implies

$$
\begin{aligned}
P_0(\widehat{G}\backslash G^*) \; \leq \; & P_0\{C_\alpha + (\Delta R_0/M_1)^{1/\underline{\gamma}} + 2T \geq r \geq C_\alpha\} \\
\leq \; & M_0\{(\Delta R_0/M_1)^{1/\underline{\gamma}} + 2T\}^{\bar{\gamma}} \, .
\end{aligned}
$$

Hence on the event $\mathcal{E}_0$,

$$
\begin{aligned}
\int_{\widehat{G}\backslash G^*} |r - C_\alpha| \, dP_0 \; \leq \; & \{(\Delta R_0/M_1)^{1/\underline{\gamma}} + 2T\} P_0(\widehat{G}\backslash G^*) \\
\leq \; & M_0\{(\Delta R_0/M_1)^{1/\underline{\gamma}} + 2T\}^{1+\bar{\gamma}} \, .
\end{aligned}
$$

We will bound $\int_{G^*\backslash\widehat{G}} |r - C_\alpha| \, dP_0$ on the event $\mathcal{E}_1 = \{R_0(\hat{\phi}) \leq \alpha\}$. Note that

$$
P_0(r \geq C_\alpha) = \alpha \geq R_0(\hat{\phi}) = P_0(\hat{r} \geq \widehat{C}_\alpha) \geq P_0(r \geq \widehat{C}_\alpha + \|\hat{r} - r\|_\infty) = P_0(r \geq \widehat{C}_\alpha + T) \, .
$$

The above chain implies that $\widehat{C}_\alpha \geq C_\alpha - T$. Therefore,

$$
\begin{aligned}
G^*\backslash\widehat{G} & = \{r < C_\alpha, \hat{r} \geq \widehat{C}_\alpha\} \\
& = \{r < C_\alpha, r \geq r - \hat{r} + \widehat{C}_\alpha\} \\
& \subset \{r < C_\alpha, r \geq \widehat{C}_\alpha - T\} \\
& \subset \{C_\alpha - 2T \leq r \leq C_\alpha\} \, .
\end{aligned}
$$

Hence on the event $\mathcal{E}_1$,

$$
\int_{G^*\backslash\widehat{G}} |r - C_\alpha| \, dP_0 \leq 2T \cdot P_0(C_\alpha - 2T \leq r \leq C_\alpha) \leq M_0(2T)^{1+\bar{\gamma}} \, ,
$$

where the last inequality follows from the margin assumption. Then it follows from (A.15) that on the event $\mathcal{E}_0 \cap \mathcal{E}_1$,

$$
\begin{aligned}
R_1(\hat{\phi}) - R_1(\phi^*) &\leq M_0 \left[\left\{ \frac{|\Delta R|}{M_1} \right\}^{1/\gamma} + 2T \right]^{1+\bar{\gamma}} + M_0(2T)^{1+\bar{\gamma}} + C_\alpha |R_0(\hat{\phi}) - R_0(\phi^*)| \\
&\leq 2M_0 \left[\left\{ \frac{\xi_{\alpha,\delta_3,m_3}(\delta_4)}{M_1} \right\}^{1/\gamma} + 2T \right]^{1+\bar{\gamma}} + C_\alpha \cdot \xi_{\alpha,\delta_3,m_3}(\delta_4) \,.
\end{aligned}
$$

From Lemma 2.1, we know that the event $\mathcal{E}_0$ occurs with probability at least $1 - \delta_4$. By Proposition 2.2 and Proposition 2.3 we know event $\mathcal{E}_1$ occurs with probability at least $1-\delta_3$, so $\mathcal{E}_0 \cap \mathcal{E}_1$ occurs with probability at least $1 - \delta_3 - \delta_4$. This completes the proof. □

Proof of Proposition 2.5. Define event

$$
\mathcal{E}_{\delta_1} = \bigcap_{j=1}^{d} \{\|\hat{F}_j^1 - F_j^1\|_\infty < \delta_1^1\} \cap \{\|\hat{F}_j^0 - F_j^0\|_\infty < \delta_1^0\} \,,
$$

where $\delta_1^1 = \sqrt{\frac{\log(4d/\delta_1)}{2n_1}}$ and $\delta_1^0 = \sqrt{\frac{\log(4d/\delta_1)}{2m_1}}$. On the event $\mathcal{E}_{\delta_1}$, for any $j \in \mathcal{A}$,

$$
\begin{aligned}
\|\hat{F}_j^0 - \hat{F}_j^1\|_\infty &\geq \|F_j^0 - F_j^1\|_\infty - \|F_j^0 - \hat{F}_j^0\|_\infty - \|F_j^1 - \hat{F}_j^1\|_\infty \\
&\geq D - \|F_j^0 - \hat{F}_j^0\|_\infty - \|F_j^1 - \hat{F}_j^1\|_\infty \\
&> D - \delta_1^0 - \delta_1^1 \,.
\end{aligned}
$$

For any $j \notin \mathcal{A}$,

$$
\begin{aligned}
\|\hat{F}_j^0 - \hat{F}_j^1\|_\infty &\leq \|F_j^0 - \hat{F}_j^0\|_\infty + \|F_j^0 - F_j^1\|_\infty + \|F_j^1 - \hat{F}_j^1\|_\infty \\
&= \|F_j^0 - \hat{F}_j^0\|_\infty + \|F_j^1 - \hat{F}_j^1\|_\infty \\
&< \delta_1^0 + \delta_1^1 \,.
\end{aligned}
$$

Since $n_1 \geq 8D^{-2}\log(4d/\delta_1)$ and $m_1 \geq 8D^{-2}\log(4d/\delta_1)$, $\delta_1^0 + \delta_1^1 \leq D - \delta_1^0 - \delta_1^1$. As a result, on the event $\mathcal{E}_{\delta_1}$, any $\tau \in \left[\delta_1^0 + \delta_1^1, D - \delta_1^0 - \delta_1^1\right]$ would lead to $\widehat{\mathcal{A}}_\tau = \mathcal{A}$. Therefore,

$$
\begin{aligned}
\mathbb{P}(\widehat{\mathcal{A}}_\tau = \mathcal{A}) &\geq \mathbb{P}(\mathcal{E}_{\delta_1}) \\
&\geq 1 - \sum_{j=1}^{d} \left\{ \mathbb{P}(\|\hat{F}_j^1 - F_j^1\|_\infty \geq \delta_1^1) + \mathbb{P}(\|\hat{F}_j^0 - F_j^0\|_\infty \geq \delta_1^0) \right\} \\
&\geq 1 - \delta_1 \,,
\end{aligned}
$$

where the last inequality follows from applying Lemma A.5 to F_j^0 and F_j^1 for $j = 1, \cdots, d$. This completes the proof. □

Proof of Proposition 2.6. Define event

$$
\mathcal{E} = \bigcap_{j \in \mathcal{A}} \{\|\log \hat{p}_j - \log p_j\|_\infty < B_j^1\} \cap \{\|\log \hat{q}_j - \log q_j\|_\infty < B_j^0\} \,,
$$

where

$$B_j^1 = \frac{C_j^1\sqrt{\frac{\log(2n_2 s/\delta_2)}{n_2 h_1}}}{\underline{\mu} - C_j^1\sqrt{\frac{\log(2n_2 s/\delta_2)}{n_2 h_1}}}, \quad B_j^0 = \frac{C_j^0\sqrt{\frac{\log(2m_2 s/\delta_2)}{m_2 h_0}}}{\underline{\mu} - C_j^0\sqrt{\frac{\log(2m_2 s/\delta_2)}{m_2 h_0}}}.$$

Let $B^1 = \sup_{j\in\mathcal{A}} B_j^1$ and $B^0 = \sup_{j\in\mathcal{A}} B_j^0$, we have $B \geq s(B^0 + B^1)$. On the event $\{\widehat{\mathcal{A}}_\tau = \mathcal{A}\} \cap \mathcal{E}$, we have

$$\log \hat{r}_{\mathrm{N}}^S(x) = \sum_{j\in\widehat{\mathcal{A}}} \log\frac{\hat{p}_j(x_j)}{\hat{q}_j(x_j)} = \sum_{j\in\mathcal{A}} \log \hat{p}_j(x_j) - \sum_{j\in\mathcal{A}} \log \hat{q}_j(x_j).$$

Therefore,

$$\begin{aligned}
\|\log \hat{r}_{\mathrm{N}}^S - \log r\|_\infty &= \|\sum_{j\in\mathcal{A}} \log\hat{p}_j - \sum_{j\in\mathcal{A}} \log\hat{q}_j - \sum_{j\in\mathcal{A}} \log p_j + \sum_{j\in\mathcal{A}} \log q_j\|_\infty \\
&\leq \sum_{j\in\mathcal{A}} (\|\log\hat{p}_j - \log p_j\|_\infty + \|\log\hat{q}_j - \log q_j\|_\infty) \\
&\leq \sum_{j\in\mathcal{A}} (B^1 + B^0) \leq B.
\end{aligned}$$

On the event $\{\widehat{\mathcal{A}}_\tau = \mathcal{A}\} \cap \mathcal{E}$, it follows from Lagrange's mean value theorem that for any x, there exists some w_x between $\log \hat{r}_{\mathrm{N}}^S(x)$ and $\log r(x)$ such that

$$\begin{aligned}
|\hat{r}_{\mathrm{N}}^S(x) - r(x)| &= |e^{\log \hat{r}_{\mathrm{N}}^S(x)} - e^{\log r(x)}| = e^{w_x}|\log \hat{r}_{\mathrm{N}}^S(x) - \log r(x)| \\
&\leq e^{\|\log r\|_\infty + B} B = Be^B\|r\|_\infty = T,
\end{aligned}$$

where the last inequality follows from the fact that

$$w_x \leq \max(\log r(x), \log \hat{r}_{\mathrm{N}}^S(x)) \leq \max(\|\log r\|_\infty, \|\log \hat{r}_{\mathrm{N}}^S\|_\infty) \leq \|\log r\|_\infty + B.$$

Thus, $\|\hat{r}_{\mathrm{N}}^S - r\|_\infty \leq T$, and we have

$$\begin{aligned}
\mathbb{P}\left(\|\hat{r}_{\mathrm{N}}^S - r\|_\infty \leq T\right) &\geq \mathbb{P}(\{\widehat{\mathcal{A}}_\tau = \mathcal{A}\} \cap \mathcal{E}) \geq \mathbb{P}(\widehat{\mathcal{A}}_\tau = \mathcal{A}) + \mathbb{P}(\mathcal{E}) - 1 \\
&= \mathbb{P}(\widehat{\mathcal{A}}_\tau = \mathcal{A}) - \mathbb{P}(\mathcal{E}^c).
\end{aligned} \tag{A.16}$$

By Proposition 2.5, we have

$$\mathbb{P}(\widehat{\mathcal{A}}_\tau = \mathcal{A}) \geq 1 - \delta_1. \tag{A.17}$$

Also, it follows from Lemma A.6 that

$$\mathbb{P}\left(\|\log\hat{p}_j - \log p_j\|_\infty > B_j^1\right) \vee \mathbb{P}\left(\|\log\hat{q}_j - \log q_j\|_\infty > B_j^0\right) \leq \delta_2/(2s).$$

Therefore,

$$\mathbb{P}(\mathcal{E}^c) \leq (2s)\delta_2/(2s) = \delta_2. \tag{A.18}$$

Plugging (A.17) and (A.18) back to (A.16) yields (2.16). Moreover, because $s \leq n_2 \wedge m_2$, it follows from Lemma A.6 that there exists some $\bar{C}_2 > 0$, such that

$$B \leq \bar{C}_2\, s\left\{\left(\frac{\log n_2}{n_2}\right)^{\frac{\beta}{2\beta+1}} + \left(\frac{\log m_2}{m_2}\right)^{\frac{\beta}{2\beta+1}}\right\}.$$

Moreover, since $s \le (n_2 \wedge m_2)^{\frac{\beta}{2(\beta+1)}}$, the above bound implies that B is bounded from above by some absolute constant. Also note that $\|r\|_\infty$ is bounded from above, so there exists an absolute constant $C_2 > 0$, such that

$$T = Be^B \|r\|_\infty \le C_2\, s \left\{ \left(\frac{\log n_2}{n_2}\right)^{\frac{\beta}{2\beta+1}} + \left(\frac{\log m_2}{m_2}\right)^{\frac{\beta}{2\beta+1}} \right\}.$$

This completes the proof. □

Proof of Theorem 2.1. Combining Propositions 2.2, 2.3, 2.4 and 2.6,

$$\mathbb{P}\left(R_0(\hat{\phi}_{\mathrm{NSN}^2}) \le \alpha, R_1(\hat{\phi}_{\mathrm{NSN}^2}) \le R_1(\phi^*) + W\right) \ge 1 - \delta_1 - \delta_2 - \delta_3 - \delta_4\,,$$

where

$$\begin{aligned} W = \;& 2M_0 \left[\left(\frac{2}{5} m_3^{1/4} M_1\right)^{-1/\gamma} + 2C_2\, s \left\{ \left(\frac{\log n_2}{n_2}\right)^{\frac{\beta}{2\beta+1}} + \left(\frac{\log m_2}{m_2}\right)^{\frac{\beta}{2\beta+1}} \right\} \right]^{1+\bar{\gamma}} \\ & + \; C_\alpha \left(\frac{2}{5} m_3^{1/4}\right)^{-1}. \end{aligned}$$

This completes the proof. □

Lemma A.5 (Dvoretzky-Kiefer-Wolfowitz inequality(Dvoretzky et al., 1956))**.** *Let X_1, X_2, $\cdots$, X_n be real-valued i.i.d. random variables with* CDF *$F(\cdot)$, and let $\hat{F}_n(x) = n^{-1}\sum_{i=1}^n \mathbb{1}(X_i \le x)$. For any $t > 0$, it holds that*

$$\mathbb{P}(\|\hat{F}_n - F\|_\infty \ge t) \;\le\; 2e^{-2nt^2}\,.$$

Or, for any given $\delta \in (0,1)$,

$$\mathbb{P}(\|\hat{F}_n - F\|_\infty \ge \sqrt{\frac{\log(2/\delta)}{2n}}) \;\le\; \delta\,. \tag{A.19}$$

Lemma A.6. *Given a density function $p \in \mathcal{P}_\Sigma(\beta, L, [-1,1])$, construct its kernel estimate $\hat{p}(x) = (nh)^{-1}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$ from i.i.d. sample $\{X_i\}_{i=1}^n$, where the kernel K is β-valid and L'-Lipschitz, and the bandwidth $h = (\log n/n)^{\frac{1}{2\beta+1}}$. For any $\delta \in (0,1)$, as long as the sample size n is such that $\sqrt{\frac{\log(n/\delta)}{nh}} < \min(1, \underline{\mu}/C)$, where $C = \sqrt{48c_1} + 32c_2 + 2Lc_3 + L' + L + \tilde{C}\sum_{1\le|l|\le\lfloor\beta\rfloor}\frac{1}{l!}$, in which $c_1 = \|p\|_\infty\|K\|_2^2$, $c_2 = \|K\|_\infty + \|p\|_\infty + \int|K||t|^\beta dt$, $c_3 = \int|K||t|^\beta dt$, and $\tilde{C}$ is such that $\tilde{C} \ge \sup_{1\le|l|\le\lfloor\beta\rfloor}\sup_{x\in[-1,1]}|p^{(l)}(x)|$, and $\underline{\mu}(>0)$ is a lower bound of p, we have*

$$\mathbb{P}\left(\|\log\hat{p} - \log p\|_\infty \ge U\right) \;\le\; \delta\,, \tag{A.20}$$

where $U = \frac{C\sqrt{\frac{\log(n/\delta)}{nh}}}{\underline{\mu} - C\sqrt{\frac{\log(n/\delta)}{nh}}}$. When $n \ge 1/\delta$, we have $U \le C_1(\log n/n)^{\frac{\beta}{2\beta+1}}$ for some absolute constant C_1.

Proof. Let $\mathcal{E}_1 = \{\|\hat{p}-p\|_\infty \le C\sqrt{\frac{\log(n/\delta)}{nh}}\}$. On the event $\mathcal{E}_1$, since $\sqrt{\frac{\log(n/\delta)}{nh}} < \min(1, \underline{\mu}/C)$, we have

$$\min(p(x_0), \hat{p}(x_0)) \ge \min(p(x_0), p(x_0) - \|\hat{p}-p\|_\infty) \ge \underline{\mu} - \|\hat{p}-p\|_\infty > 0\,.$$

It then follows from Lagrange's mean value theorem that for any fixed x_0, there exists some w_{x_0} between $\hat{p}(x_0)$ and $p(x_0)$,

$$\begin{aligned} |\log\hat{p}(x_0) - \log p(x_0)| &= w_{x_0}^{-1}\,|\hat{p}(x_0) - p(x_0)| \\ &\le [\min\{\hat{p}(x_0), p(x_0)\}]^{-1}|\hat{p}(x_0) - p(x_0)| \le \frac{\|\hat{p}-p\|_\infty}{\underline{\mu} - \|\hat{p}-p\|_\infty}. \end{aligned}$$

As a result, it holds on event $\mathcal{E}_1$ that

$$\|\log\hat{p} - \log p\|_\infty \;\le\; \frac{C\sqrt{\frac{\log(n/\delta)}{nh}}}{\underline{\mu} - C\sqrt{\frac{\log(n/\delta)}{nh}}} \;=\; U\,,$$

and

$$\mathrm{I\!P}(\|\log\hat{p} - \log p\|_\infty \le U) \ge \mathrm{I\!P}(\|\hat{p}-p\|_\infty \le C\sqrt{\frac{\log(n/\delta)}{nh}}) \ge 1-\delta\,,$$

where the last inequality follows from Lemma A.1 in Tong (2013) (the special case of $d=1$). Finally when $n \ge 1/\delta$, we have $U = \frac{C\sqrt{\frac{\log(n/\delta)}{nh}}}{\underline{\mu}-C\sqrt{\frac{\log(n/\delta)}{nh}}} \le C_1\,(\log n/n)^{\frac{\beta}{2\beta+1}}$ for some absolute constant C_1. This completes the proof. □

Appendix B. About detection condition and Assumption 3

We show that it is possible for densities satisfying Assumption 3 to violate a generalized version of the detection condition defined in Definition 2.3. While the generalized detection condition applies to general (P, f, C^*) as the original one, we narrow its definition to (P_0, r, C_α) which we actually use in the main text.

Definition B.1 (Generalized detection condition)**.** *Let $u(\cdot)$ be a strictly increasing differentiable function on $\mathbb{R}^+$ with $\lim_{x\to 0+} u(x) = 0$, a function $r(\cdot)$ is said to satisfy the* generalized detection condition *with respect to P_0 and $u(\cdot)$ at level (C_α, δ^*) if for any $\delta \in (0, \delta^*)$,*

$$P_0\{C_\alpha \le r(X) \le C_\alpha + \delta\} \quad\ge\quad u(\delta)\,. \tag{B.1}$$

The following conditions suffice to make (B.1) fail

$$P_0\left\{C_\alpha \;\le\; r(X) \;\le\; C_\alpha + k^{-1}\right\} \quad<\quad u(k^{-1})\,,\quad k = 1, 2, \ldots\,. \tag{B.2}$$

A 1-dimensional toy example that satisfies Assumption 3 and (B.2) (thus violating the generalized detection condition) is given as follows. Assume P_0 and P_1 have the same

support $[-1,1]$. Given $u(\cdot)$ as a strictly increasing differentiable function on $\mathbb{R}^+$ with $\lim_{x\to 0+} u(x) = 0$, let $q(x) = \alpha$ for all $x \in [0,1]$, and set $p(x)$ accordingly such that

$$r(x) \;=\; \frac{p(x)}{q(x)} \;=\; \begin{cases} 2u^{-1}(1) + 2u^{-1}(\alpha x)\,, & x \in (0,1]\,, \\ 2u^{-1}(1)\,, & x = 0\,, \\ 2u^{-1}(1) - v(x)\,, & x \in [-1,0)\,, \end{cases} \tag{B.3}$$

where $v(\cdot)$ is some positive differentiable function that makes $r(\cdot)$ differentiable at $x = 0$. It follows from (B.3) that $\left\{x \in [-1,1] : r(x) \geq 2u^{-1}(1)\right\} = [0,1]$, and identity

$$P_0\left\{r(X) \geq 2u^{-1}(1)\right\} \;=\; \int_{\{x\in[-1,1]:\, r(x)\geq 2u^{-1}(1)\}} q(x)dx \;=\; \int_{[0,1]} q(x)dx \;=\; \alpha$$

implies $C_\alpha = 2u^{-1}(1)$. As a result, for any $k \in \{1,2,\ldots\}$ we have

$$\left\{C_\alpha \leq r(X) \leq C_\alpha + k^{-1}\right\} \;=\; \left\{X \in [0,1],\, 2u^{-1}(\alpha X) \leq k^{-1}\right\} \;=\; \left\{X \in \left[0, \alpha^{-1}u(0.5k^{-1})\right]\right\},$$

and

$$\begin{aligned} P_0\left\{C_\alpha \,\leq\, r(X) \,\leq\, C_\alpha + k^{-1}\right\} \;&=\; P_0\left\{X \in \left[0, \alpha^{-1}u(0.5k^{-1})\right]\right\} \;=\; \int_0^{\alpha^{-1}u(0.5k^{-1})} q(x)dx \\ &=\; \alpha\cdot\alpha^{-1}u(0.5k^{-1}) \;=\; u\left(0.5k^{-1}\right) \;<\; u\left(k^{-1}\right) \end{aligned}$$

satisfies (B.2). Note that the above construction makes no assumption about the behavior of $q(\cdot)$ and $p(\cdot)$ on $[-1,0)$ except the normalization constraints $\int_{[-1,1]} pdx = \int_{[-1,1]} qdx = 1$ and $r(\cdot)$ being differentiable on $[-1,1]$. Thus, there exist p, q, and r that satisfy Assumption 3.

Appendix C. An alternative threshold estimate

This part contains an alternative estimate of threshold C_α that guarantees type I error bound. Based on Chernoff inequality, the following Proposition gives an alternative version of Proposition 2.2. First, we introduce two technical lemmas.

Lemma C.1. *If $G_k \sim$ Gamma$(k,1)$, $k > 0$, then for any $\tau \in (0,k)$, we have*

$$\mathbb{P}(G_k \geq k+\tau) \;\leq\; e^{-\tau^2/(4k)}, \quad \mathbb{P}(G_k \leq k-\tau) \;\leq\; e^{-\tau^2/(2k)} \;\leq\; e^{-\tau^2/(4k)}.$$

Proof of Lemma C.1. For any $\epsilon \in (0,1)$ and $t \in (0,1)$, it follows from Chernoff inequality that

$$\mathbb{P}\left\{G_k \geq (1+\epsilon)k\right\} \;=\; \mathbb{P}\left\{e^{tG_k} \geq e^{t(1+\epsilon)k}\right\} \;\leq\; \frac{\mathbb{E}(e^{tG_k})}{e^{t(1+\epsilon)k}} \;=\; (1-t)^{-k}e^{-t(1+\epsilon)k}. \tag{C.1}$$

Letting $t = \operatorname{argmin}_{x\in(0,1)}(1-x)^{-k}e^{-x(1+\epsilon)k} = \epsilon/(1+\epsilon)$ in (C.1) yields

$$\mathbb{P}\left\{G_k \geq (1+\epsilon)k\right\} \;\leq\; (1+\epsilon)^k e^{-\epsilon k} \;=\; e^{k\{\log(1+\epsilon)-\epsilon\}} \;\leq\; e^{-k\epsilon^2/4},$$

Likewise, for any $\epsilon \in (0,1)$ and $s<0$,

$$\mathbb{P}\left\{G_k \le (1-\epsilon)k\right\} \;=\; \mathbb{P}\left\{e^{sG_k} \ge e^{s(1-\epsilon)k}\right\} \;=\; (1-s)^{-k}e^{-s(1-\epsilon)k}. \tag{C.2}$$

Letting $s = \operatorname{argmin}_{x<0}(1-x)^{-k}e^{-x(1-\epsilon)k} = -\epsilon/(1-\epsilon)$ in (C.2) yields

$$\mathbb{P}\left\{G_k \le (1-\epsilon)k\right\} \;\le\; (1-\epsilon)^k e^{\epsilon k} \;=\; e^{k\{\log(1-\epsilon)+\epsilon\}} \;\le\; e^{-k\epsilon^2/2},$$

where the last inequality follows from Taylor expansion

$$\log(1-\epsilon)+\epsilon \;=\; \sum_{i=1}^{\infty}\frac{\epsilon^i}{i}-\epsilon \;=\; \sum_{i=2}^{\infty}\frac{\epsilon^i}{i} \;>\; \frac{\epsilon^2}{2}, \quad \forall 0<\epsilon<1\,.$$

Take $\epsilon = \tau/k$, the conclusion of the lemma follows.

□

Lemma C.2. *Let $B \sim Beta(a,b)$, and $\mu = \mathbb{E}(B) = a/(a+b)$. For any $t \in (0, 1-\mu)$,*

$$\mathbb{P}\left\{B > \mu+t\right\} \;\le\; 2\exp\left[-4^{-1}\left\{\frac{(a+b)t}{\sqrt{b}(\mu+t)+\sqrt{a}(1-\mu-t)}\right\}^2\right].$$

Proof of Lemma C.2. By properties of beta distribution, we can represent B as

$$B \;=\; \frac{G_a}{G_a+G_b}, \quad \text{where } G_a \sim \Gamma(a,1),\ G_b \sim \Gamma(b,1) \text{ are independent.}$$

For any $t>0$ and constant C such that $a(1-\mu-t) < C < b(\mu+t)$, we have

$$\begin{aligned}
\mathbb{P}(B \le \mu+t) \;&=\; \mathbb{P}\left\{(1-\mu-t)G_a \le (\mu+t)G_b\right\} \;\ge\; \mathbb{P}\left\{(1-\mu-t)G_a \le C \le (\mu+t)G_b\right\} \\
&=\; \mathbb{P}\left\{(1-\mu-t)G_a \le C\right\}\mathbb{P}\left\{C \le (\mu+t)G_b\right\} \\
&=\; \mathbb{P}\left(G_a \le \frac{C}{1-\mu-t}\right)\mathbb{P}\left(G_b \ge \frac{C}{\mu+t}\right) \\
&=\; \left\{1-\mathbb{P}\left(G_a > \frac{C}{1-\mu-t}\right)\right\}\left\{1-\mathbb{P}\left(G_b > \frac{C}{\mu+t}\right)\right\} \\
&\ge\; 1-\mathbb{P}\left(G_a > \frac{C}{1-\mu-t}\right)-\mathbb{P}\left(G_b > \frac{C}{\mu+t}\right),
\end{aligned} \tag{C.3}$$

where by Lemma C.1

$$\begin{aligned}
\mathbb{P}\left(G_a > \frac{C}{1-\mu-t}\right) \;&\le\; \mathbb{P}\left\{G_a > a + \left(\frac{C}{1-\mu-t}-a\right)\right\} \;\le\; e^{-\left(\frac{C}{1-\mu-t}-a\right)^2(4a)^{-1}}, \\
\mathbb{P}\left(G_b < \frac{C}{\mu+t}\right) \;&\le\; \mathbb{P}\left\{G_b < b - \left(b-\frac{C}{\mu+t}\right)\right\} \;\le\; e^{-\left(b-\frac{C}{\mu+t}\right)^2(4b)^{-1}}.
\end{aligned} \tag{C.4}$$

Letting

$$C \;=\; \frac{(1-\mu-t)(\mu+t)(a\sqrt{b}+\sqrt{a}b)}{\sqrt{b}(\mu+t)+\sqrt{a}(1-\mu-t)}$$

in (C.3) such that the two exponents in (C.4) equal

$$\left(\frac{C}{1-\mu-t}-a\right)^2(4a)^{-1} \;=\; \left(b-\frac{C}{\mu+t}\right)^2(4b)^{-1} \;=\; 4^{-1}\left\{\frac{(a+b)t}{\sqrt{b}(\mu+t)+\sqrt{a}(1-\mu-t)}\right\}^2$$

yields

$$\begin{aligned}
\mathbb{P}(B>\mu+t) &= 1-\mathbb{P}(B\le\mu+t) \;\le\; \mathbb{P}\left(G_a>\frac{C}{1-\mu-t}\right)+\mathbb{P}\left(G_b>\frac{C}{\mu+t}\right)\\
&\le e^{-\left(\frac{C}{1-\mu-t}-a\right)^2(4a)^{-1}}+e^{-\left(b-\frac{C}{\mu+t}\right)^2(4b)^{-1}}\\
&= 2\exp\left[-4^{-1}\left\{\frac{(a+b)t}{\sqrt{b}(\mu+t)+\sqrt{a}(1-\mu-t)}\right\}^2\right].
\end{aligned}$$

This completes the proof. □

Proposition C.1. *Let $\hat{r}(\cdot)$ be any estimate of the density ratio function. For any $\delta_3\in(0,1)$ and $k\in\{1,\cdots,m_3\}$, the type I error of classifier $\hat{\phi}_k$ defined in (2.1) satisfies*

$$\mathbb{P}\left\{R_0(\hat{\phi}_k)>h(\delta_3,m_3,k)\right\} \;\le\; \delta_3\,,$$

where

$$h(\delta_3,m_3,k)=\frac{m_3+1-k+2\sqrt{\log(2/\delta_3)}\sqrt{m_3-k+1}}{m_3+1+2\sqrt{\log(2/\delta_3)}\left(\sqrt{m_3-k+1}-\sqrt{k}\right)}\,.$$

Proof. Let B be a realization from Beta(k,m_3+1-k). It follows from Proposition 2.1 that

$$\begin{aligned}
\mathbb{P}\{R_0(\hat{\phi}_k)>h(\delta_3,m_3,k)\} &\le \text{Beta.cdf}_{k,m_3+1-k}\{1-h(\delta_3,m_3,k)\}\\
&= \mathbb{P}\{B\le 1-h(\delta_3,m_3,k)\} \;=\; \mathbb{P}\{1-B\ge h(\delta_3,m_3,k)\}
\end{aligned}$$

for any $k\in\{1,\cdots,m_3\}$ and $\hat{r}$, with $1-B\sim$ Beta(m_3+1-k,k). Letting $a=m_3+1-k$, $b=k$, and

$$t=\frac{2\sqrt{\log(2/\delta_3)}\left\{(m_3+1-k)\sqrt{k}+k\sqrt{m_3+1-k}\right\}}{(m_3+1)\left\{m_3+1+2\sqrt{\log(2/\delta_3)}\left(\sqrt{m_3+1-k}-\sqrt{k}\right)\right\}}\,.$$

in Lemma C.2 yields

$$\mathbb{P}\left\{R_0(\hat{\phi}_k)>h(\delta_3,m_3,k)\right\} \;\le\; \delta_3\,.$$

This completes the proof. □

Proposition C.1 implies that $h(\delta_3,m_3,k)\le\alpha$ is a sufficient condition for the classifier $\hat{\phi}_k$ (defined in (2.2)) to satisfy NP Oracle Inequality (I) ($k=1,\ldots,m_3$). Let $\mathcal{K}_{\text{chern}}=\{k\in\{1,\cdots,m_3\}:h(\delta_3,m_3,k)\le\alpha\}$. Similar to Proposition 2.3 we can prove $\mathcal{K}_{\text{chern}}$ to be non-empty as long as m_3 is greater than some threshold.

Numerical investigation shows that for most combinations of (α, δ_3, m_3) with non-empty $\mathcal{K}$ and $\mathcal{K}_{\text{chern}}$, $k_{\min} = \min_k \mathcal{K}$ as defined in (2.7) is better than $k_{\text{chern}} = \min_k \mathcal{K}_{\text{chern}}$ in the sense that $\hat{\phi}_{k_{\min}}$ has a lower type II error than $\hat{\phi}_{k_{\text{chern}}}$ as a result of $k_{\min} < k_{\text{chern}}$. Specifically, for each $\delta_3 \in \{0.01 \cdot i\}_{i=1}^{10}$, the number of $\{k_{\text{chern}} < k_{\min}\}$ out of 100 combinations of $(\alpha, m_3) \in \{0.01 \cdot i\}_{i=1}^{10} \times \{100 \cdot i\}_{i=1}^{10}$ is reported as follows. Only when δ_3 gets very close to 0 is k_{chern} preferred to $k_{\min}$.

δ_3	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
$\#\{k_{\text{chern}} < k_{\min}\}$	83	70	49	4	0	0	0	0	0	0

References

Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10** 1471–2105.

Alderton, G. K. (2014). Breast cancer: Breast cancer classification. *Nature Reviews Cancer*, **14** 155–155.

Audibert, J. and Tsybakov, A. (2007). Fast learning rates for plug-in classifiers under the margin condition. *The Annals of Statistics*, **35** 608–633.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10** 989–1010.

Blanchard, G., Lee, G. and Scott, C. (2010). Semi-supervised novelty detection. *Journal of Machine Learning Research*, **11** 2973–3009.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, **106** 1566–1577.

Cannon, A., Howse, J., Hush, D. and Scovel, C. (2002). Learning with the neyman-pearson and min-max criteria. *Technical Report LA-UR-02-2951.*

Casasent, D. and Chen, X. (2003). Radial basis function neural networks for nonlinear fisher discrimination and neyman-pearson classification. *Neural Networks*, **16** 529 – 535.

Danziger, S. A., Swamidass, S. J., Zeng, J., Dearth, L. R., Lu, Q., Chen, J. H., Cheng, J., Hoang, V. P., Saigo, H., Luo, R. et al. (2006). Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **3** 114–125.

Dümbgen, L., Igl, B. and Munk, A. (2008). P-value for classification. *Electronic Journal of Statistics*, **2** 468–493.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics* 642–669.

Elkan, C. (2001). The foundations of cost-sensitive learning. *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* 973–978.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, **36** 2605–2637.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, **106**.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74** 745–771.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70** 849–911.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286** 531–537.

Hall, P., Pham, T. et al. (2010). Optimal properties of centroid-based classifiers for very high-dimensional data. *The Annals of Statistics*, **38** 1071–1093.

Han, M., Chen, D. and Sun, Z. (2008). Analysis to Neyman-Pearson classification with convex loss function. *Analysis in Theory and Applications*, **24** 18–28.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer-Verlag Inc.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Koltchinskii, V. (2008). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.*

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107** 1129–1139.

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, **99** 29–42.

Mammen, E. and Tsybakov, A. (1999). Smooth discrimination analysis. *The Annals of Statistics*, **27** 1808–1829.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72** 417–473.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Knig, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006). Customized oligonucleotide microarray gene expressionbased classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology*, **24** 5070–5078.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, **23** 855–881.

Rigollet, P. and Tong, X. (2011). Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, **12** 2831–2855.

Robins, J., Van Der Vaart, A. et al. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, **34** 229–253.

Scott, C. (2005). Comparison and design of neyman-pearson classifiers. Unpublished.

Scott, C. (2007). Performance measures for Neyman-Pearson classification. *IEEE Transactions on Information Theory*, **53** 2852–2863.

Scott, C. and Nowak, R. (2005). A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, **51** 3806–3819.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, **39** 1241–1265.

Tarigan, B. and van de Geer, S. (2006). Classifiers of support vector machine type with l1 complexity regularization. *Bernoulli*, **12** 1045–1076.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **99** 6567–6572.

Tong, X. (2013). A plug-in approach to nayman-pearson classification. *Journal of Machine Learning Research*, **14** 3011–3040.

Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, **32** 135–166.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer.

Tsybakov, A. and van de Geer, S. (2005). Square root penalty: Adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, **33** 1203–1224.

Weiss, S. M., Indurkhya, N., Zhang, T. and Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer.

Witten, D. and Tibshirani, R. (2012). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society Series B*, **73** 753–772.

Wu, S., Lin, K., Chen, C. and M., C. (2008). *Asymmetric support vector machines: low false-positive learning under the user tolerance.*

Yang, Y. (1999). Minimax nonparametric classification-part i: rates of convergence. *IEEE Transaction Information Theory*, **45** 2271–2284.

Zadrozny, B., Langford, J. and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. *IEEE International Conference on Data Mining* 435.